EVIDENCE ABOUT THE CONVERGENT AND DISCRIMINANT VALIDITY OF THE COMPONENTS OF FIDELITY OF IMPLEMENTATION IN A SCIENCE-TEACHER PROFESSIONAL DEVELOPMENT PROJECT

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

EDUCATIONAL PSYCHOLOGY

DECEMBER 2017

By

Brian E. Lawton

Dissertation Committee:

Paul R. Brandon, Chairperson
Lois A. Yamauchi
Anna Ah Sam
George M. Harrison
Ronald H. Heck

Keywords: fidelity of implementation, multitrait-multimethod matrix, convergent validity, discriminant validity, inquiry-based science professional development

ACKNOWLEDGEMENTS

I would like to thank the members of my dissertation committee for their continued support during my lengthy dissertation process. I especially would like to thank Paul Brandon, my chairperson and mentor. I am forever grateful for his enduring guidance and encouragement through this journey.

I would like to thank the TSI-A project development and evaluation team members whose patience and flexibility supported my dissertation efforts every step of the way. I especially would like to thank Lisa Vallin for all her time with helping me with my observations.

I would like to acknowledge my daughters, Isabelle and Ruby, who are always on my mind. I would also like to thank Angel for feeding me and providing emotional support during my long hours in front of the computer.

Finally, my greatest thanks go to my parents, Rick and Linda. Without them none of this would have been possible.

ABSTRACT

Many evaluation theorists, methodologists, and practitioners have concluded that measuring fidelity of implementation is an essential element in program evaluation studies. Evaluators must not only be able to present evidence of a program's effectiveness but also must present evidence about how a program worked. Research about fidelity of implementation to date has mostly focused on identifying, defining, and determining how best to measure fidelity of implementation. However, the research has not fully explored the relationships among the different components of fidelity of implementation and the extent to which they are distinct from each other. Understanding these relationships has implications for how we conceptualize fidelity of implementation, what needs to be considered when developing fidelity of implementation measures, and ultimately what components of fidelity of implementation are the best predictors of positive study outcomes. The goal of this correlational study was to provide a systematic examination of the interrelationships among four of the primary components of fidelity of implementation. Fidelity of implementation data were collected from instruments used to evaluate a small-scale inquiry-based science professional development project. The data were organized into a multitrait-multimethod matrix to examine the convergent and discriminant validity of each of the components. The results provide mixed evidence about the relationships among, and distinctness of, the components. The findings suggest that the methods of measurement, the extent to which the components were represented on multiple instruments, and the level of the teachers' understanding of the components affected the extent to which convergent and discriminant validity could be shown. The findings contribute to the research and evaluation literature by providing insight into underreported components of fidelity of implementation, their relationships to each other, and what researchers and evaluators need to consider when collecting fidelity of implementation data.

CONTENTS

TABLES

FIGURES

CHAPTER 1

INTRODUCTION

Measuring fidelity of implementation (FOI) is of great importance to evaluators studying the extent to which a program functions as intended. Without measures of FOI during program implementation, it may be unclear whether unsuccessful outcomes reflect a failure of the program or a failure to implement the program as intended. Studies that include measures of FOI increase the validity and feasibility of an intervention, protect against inaccurate conclusions about a program's effectiveness (e.g., Type III errors), and provide information about how the program might be improved in future interventions (Dusenbury et al., 2003; Ruiz-Primo, 2006; Sanchez et al., 2007). Conversely, when programs are not implemented with fidelity, they may be less effective, efficient, and predictable (Noell, Gresham, & Gansle, 2002; Wilder, Atwell, & Wine, 2006). Additionally, measuring FOI helps evaluators document and recommend changes to an intervention (Lane, Bocian, Macmillan, & Gresham, 2004); helps evaluators understand the limits of the interventions, as well as their generalizability to other populations and settings (LeLaurin & Wolery, 1992); and informs replication efforts (Gresham, Gansle, & Noell, 1993; Gresham, MacMillan, Beebe-Frankenberger, & Bocian, 2000; Lane et al., 2004; Moncher & Prinz, 1991).

FOI has seen increased attention over the past decade, as evident by the numerous review studies on the topic (for example, see Durlak & Dupre, 2008; Dusenbury et al., 2003; Fixsen, Naoom, Blasé, Friedman, & Wallace, 2005; Meyers & Brandt, 2014; Mowbray, Holter, Teague, & Bybee, 2003; O'Donnell, 2008). However, creating a standard for studying FOI has posed several challenges to evaluators. These include (a) limited agreement about how best to define FOI, (b) a lack of consensus about how best to measure FOI, and (c) inconsistent results about the extent to which FOI is related to study outcomes. Therefore, much of the current research to

date has worked on better identifying and defining the constructs that make up FOI, improving the way FOI data are collected, and determining the extent to which FOI positively effects outcomes. However, research has not fully explored the relationships among the different components of FOI and the extent to which they are distinct. Understanding the relationships between and among the FOI components has implications for how we define FOI and how best to measure different components of FOI, and ultimately it will help determine which components of FOI are the best predictors of positive study outcomes.

Statement of Purpose

The purpose of my study is to examine the *convergent* and *discriminant* validity of the FOI components as they are operationalized in a small-scale inquiry-based science professional development project. My study will examine the major components of FOI discussed in the educational and evaluation literature, including adherence, exposure, quality, and participant responsiveness. I have drawn upon data collected within the context of a three-year professional development (PD) project evaluation study. The purpose of the PD project was to enhance elementary, middle, and high school teachers' inquiry-based science teaching using aquatic science content. The evaluators collected data, using multiple instruments, to examine the extent to which the PD had an effect on the participating teachers' inquiry-based instruction skills, aquatic science content knowledge, and their pedagogical knowledge when using inquiry-based science techniques (Seraphin, 2014). My study includes data collected from three of the instruments that were administered to collect FOI and other process data during the final year of the project. I also used an observation protocol designed to measure inquiry-based science instruction practices to gather additional FOI data on project teachers.

Research and evaluation theory has suggested that there are multiple components comprising FOI. My intent is to examine this theory using data collected about FOI components as they were

operationalized in the PD study. Toward this end, I use Campbell and Fiske's (1959) multitrait-multimethod (MTMM) approach for examining convergent and discriminant validity.

Campbell and Fiske developed the MTMM approach for determining the validity about specific instruments as they are used to measure specific traits. In my study, however, the primary focus is not on the implications of MTMM findings for the validity of inferences from specific instruments. Instead, I focus on the broad implications of a *specific* MTMM study for researchers' and evaluators' *general* understanding of the relationships among FOI components. This is not to say that the findings of my study are inapplicable to the instruments that I examine, but such an application is ancillary to my purposes. My purposes are to (a) gain insights into the theory of FOI, (b) provide new information about the degree to which FOI components are differentiated, (c) increase our understanding of how FOI components might be conceptualized in future research and evaluation studies, and (d) contribute to the research and evaluation FOI literature by providing suggestions about how the components might best be measured.

Research Questions

I have two research questions that guide my study:

1. To what extent do the components of FOI as measured by instruments used in the TSI-A evaluation demonstrate convergent validity?
2. To what extent do the components of FOI as measured by instruments used in the TSI-A evaluation demonstrate discriminant validity?

Definition of Key Terms

To provide a basic foundation of the terminology that is used throughout my study, I briefly define the relevant key terms in this section. I will expand upon the description of these key terms throughout this study to ensure clarity and understanding. The definitions are presented in alphabetic order.

Adherence

Adherence is the extent to which specified program components are delivered in the manner as the program prescribed (Dane & Schneider, 1998). For my study, I defined adherence as the extent to which participants addressed and implemented the lessons and activities in accordance with the program guidelines (e.g., the extent to which all the steps in the program's required target activities were implemented).

Exposure

Exposure refers to the amount of the program content that is received by participants, as well as the frequency in which specific program techniques are implemented (Dane & Scheider, 1998). For my study, exposure is defined as how often the teachers addressed each of the project components.

Convergent Validity

Convergent validity is the degree to which concepts that should be related theoretically are related in reality (Campbell & Fiske, 1959).

Discriminant Validity

Discriminant validity is the degree to which concepts that should *not* be related theoretically are, in fact, *not r*elated in reality (Campbell & Fiske, 1959).

Participant Responsiveness

Participant responsiveness is the participant's level of participation in the intervention, the degree of interest in or the perceived relevance of the program by the participant, and extent to which the participants' are engaged in the program's activities (Carroll et al., 2007; Durlak & DuPre, 2008; Dusenbury et al., 2003; Lynch & O'Donnell, 2005). For my study, participant responsiveness is defined as the extent to which the teachers (a) found value in the project, including the extent to which it enhanced their understanding of a program components and

science topics; (b) perceived their implementation to be successful; and (c) perceived the project to positively affect student learning and engagement.

Quality

Quality looks at how well the participant implements the program's components in the specified setting (Dusenbury, Brannigan, Falco, & Hansen, 2003; Ruiz-Primo, 2006). For my study, quality is defined as how well the teachers implemented different components of an activity in accordance with program theory.

CHAPTER 2

REVIEW OF THE LITERATURE

This review of the literature provides the foundation for my examination of the relationship between the commonly discussed components of FOI. I provide an overview of the history and definition of FOI; describe how the components are represented in the literature; describe why FOI is an important measure in outcome studies, including the different methods of measuring FOI. I conclude with an overview of how I will examine the FOI components central to my study.

Understanding the relationship between a program's design and the extent to which it is implemented is of great concern to the developers and evaluators held accountable for determining what aspects of a program influence intended outcomes. Evaluators and practitioners can gain a better understanding of how and why an intervention works by measuring whether it has been implemented as intended. It has become increasingly common for evaluators to examine the implementation of program practices—providing the evidence not only of a program's successes and failures but also what factors produced the effects. Unless such an examination occurs, the degree of impact cannot be associated with the level and extent of program implementation (Carroll et al., 2007). Measuring FOI also allows for greater confidence in the results by providing enhanced statistical evidence that program components are delivered consistently across participants (e.g., individuals or classrooms) and that the implementation is true to the program model and theory (Dumas, Lynch, Laughlin, Smith, & Prinz, 2001; Teague, Drake, & Ackerson, 1997). Ensuring that a program is being used as intended, or implemented with fidelity, is an often under-reported or even overlooked aspect of an evaluation (Brandon, Taum, Young, Pottenger, & Speitel, 2008; Mowbray, et al., 2003; O'Donnell, 2008)—arguably, the aspect of an evaluation that provides the documentation and

support for interpreting the evaluation findings. Therefore, collecting fidelity data provides an in-depth understanding of program implementation, including program progress and needed revisions; a way to examine theoretical assumptions; improved interpretation of outcomes and relationships among variables; and a means of offering timely feedback to both participants and developers for program improvement. (Backer, 2001; Dane & Schneider, 1998; Domitrovich & Greenberg, 2000; Mowbray et al., 2003; Pankratz et al., 2006). FOI has gained increased attention as a mediating variable on the outcomes of educational interventions (Dane & Schneider, 1998; Mowbray, et al., 2003; O'Donnell, 2008). Importantly, including fidelity measures also promotes validity, both internal and external, by providing documentation, guidelines, and criteria for replicating programs (Cook & Campbell, 1979; Dumas et al., 2001; Dusenbury et al., 2003; Mowbray et al., 2003; O'Donnell, 2008).

FOI measurements occur during efficacy and effectiveness studies to ensure FOI at various stages (Mills & Ragan, 2000; O'Donnell, 2008). During an efficacy study, measuring FOI is a process completed to ensure internal validity. The process involves constant monitoring to isolate the critical components of the program and to make revisions based on what was occurring during the implementation for each phase of the study (Resnick et al., 2005; Mowbray et al., 2003). In effectiveness studies, FOI is used to determine the extent to which the intervention produces a desired effect when the program is actually used (Sanchez et al., 2007). Without measuring FOI, it is nearly impossible for research to determine the extent to which either successful or unsuccessful outcomes are the result the program model or to the implementation of the program model. According to O'Donnell (2008) and Greenberg, Domitrovich, Graczyk, & Zins (2005) effectiveness studies can determine the capacity for scaling up a program. My study is occurring within the context of a larger effectiveness study

that was implemented in a school setting to examine the effects of the program on several student and teacher outcome measures. The ultimate goal of my study is to determine the extent to which components of FOI, as operationalized in this study, are valid and distinct from each other.

Brief History of Fidelity of Implementation

The concept of FOI has been around for some time. It was identified in early rural sociology of the 1920s and 30s in the study of farmers' use and adaptation of new developing technologies of the time (e.g., see Ryan & Gross, 1943). This early research led to the development of the *diffusion of innovation* theory (Rogers, 1962). Dusenbury et al. (2003) described diffusion of innovation as "a way of understanding the process by which new ideas are put into practice" (p. 238). It was this foundation that led to the development of the *research, development, and diffusion* (RD&D) model, which emphasized the need to perform rigorous research in demonstration projects. The basic assumptions of this model are that a well-researched, effective, and validated program will produce positive results; users will adopt a program that has shown to have positive results; and once adopted, users will implement the program consistent with the model across sites and conditions (Dusenbury et al., 2003; Emshoff et al., 1987). Toward the end of the 1970s, evaluators and researchers began to question the assumption that educators, social-service providers, and others implementing programs are viewed as largely passive and that they would implement a program as intended by developers simply based on positive results. Researchers argued that it is not just the positive results of a program that may influence implementation but rather the characteristics of the individual organizations, its members, and the setting in which it is applied that affect the extent to which a program is implemented with fidelity, or adopted at all (Berman & McLaughlin, 1976; Dusenbury et al., 2003; Fixsen et al., 2005). Therefore, as Lynch (2007) suggests, evaluators must not only present evidence of a program's effectiveness but also must present evidence about how and why it worked. Berman &

McLaughlin (1976) underscored the importance of assessing program fidelity, arguing that to produce change, an intervention must be implemented with fidelity, and without an assessment of fidelity, there is no way to determine whether unsuccessful outcomes reflect a failure of the model or failure to implement the model as intended (i.e., Type III error).

The importance of implementation fidelity became most evident in early psychotherapy research. Because of the minimal description of the treatments prescribed it was difficult to replicate effective interventions (Bond, Evans, Salyers, Williams, & Kim, 2000). As a result of this lack of description, it became critical to ensure that effective treatments described the central components of an intervention, as well as the modes in which the intervention was implemented. As Bond et al. (2000) noted, "Fidelity measurement accelerated the maturation of psychotherapy research by making standardized treatments possible and by providing methods to document differences between different forms of treatment" (p. 76). However, even with the obvious benefits of greater documentation needed for the replication of effective treatments, Moncher and Prinz (1991) found that research and evaluation studies only marginally described programs' adherence to protocol or only used some form of a standard treatment manual as evidence of adherence. These findings are consistent with Dane & Schneider's (1998) conclusions in their review of primary and early secondary prevention programs between 1980 and 1994. They found that less than a quarter of the outcome studies reviewed specified procedures for the documentation of fidelity. Similarly, educational research and evaluation also suffered from this deficiency by failing to provide adequate descriptions of the procedures and central features of treatments that were shown to produce positive consumer effects (Dhillon, Darrow, & Meyers, 2015)—descriptions that are critical to validating the findings and disseminating effective programs. Therefore, even with the increased interest and clear importance of studying FOI over

the last several decades, it has yet to become fully incorporated in much of the research and evaluation.

FOI Defined

There is a lack of consensus in the broader implementation literature (i.e., public health, criminal justice, education, preventative medicine) for common language and operational definitions of FOI terminology (Century & Cassata, 2016; Durlak & DuPre, 2008; O'Donnell, 2008). However, FOI in the field of educational evaluation and research has largely been defined as *the degree to which a teacher or other program provider implements a program as originally intended by the program developers* (Dane and Schneider 1998; Dusenbury et al., 2003; Lynch, 2007; Lynch & O'Donnell, 2005; Mowbray et al., 2003; Ruiz-Primo, 2006). Therefore, this is the general definition that I use for this study.

FOI Components

The variation in how participants receive an intervention and the degree to which it is delivered informs how effective a program might be across settings. An intervention can be delivered with a high degree of skill and integrity, but participants still may not receive or interact with the intervention as intended. Breakdowns occur when participants are not engaged during treatment delivery, fail to comprehend or follow through on intervention protocols, and/or only intermittently attend sessions (Zvoch, 2012). Therefore, to address these issues, researchers began to identify the various components to get a more accurate picture of FOI. The most commonly discussed components of FOI in the literature have been (a) *adherence*—the congruence between program delivery and program design, (b) *exposure* or *dose*—the frequency or duration of program activities, (c) *quality*—how well a program is implemented using the methods prescribed, (d) *participant responsiveness*—the extent of program participants' positive involvement in program activities, and (e) *program differentiation*—the degree to which a

program is distinguishable between comparison conditions or competing programs that have the same theoretical background (Berkel, Mauricio, Schoenfelder, & Sandler, 2011, Dane & Schneider 1998; Durlak & DuPre, 2008; Dusenbury et al., 2003; Giles et al., 2008; Hill & Owens, 2013; Ibrahim & Sidani, 2016).

For my study, adherence is defined as the extent to which the teachers addressed and implemented the lessons and activities in accordance with the program guidelines (e.g., the extent to which all the steps in the target activities were implemented). Exposure is defined as the amount or frequency that the teachers addressed the project components during implementation. Quality is defined as how well the teachers implemented the different components of the project in accordance with the project theory. Participant responsiveness is defined as the extent to which the teachers found value in the program and the teachers' perceived effectiveness of an activity on student learning. Because data were not collected to identify the presence of critical program components that may have been present outside of the intervention, I did not include program differentiation in my study. Although program differentiation is commonly included as a primary component of FOI, Century, Rudnick, and Freeman (2010) posit that program differentiation is more of a measure about the extent that there are shared critical components between programs (i.e., that there is a distinct identifiable intervention when compared to a comparison condition or other treatments) rather than a measure of fidelity. Also, Darrow (2013) notes program differentiation is less of a factor once the program's core components have been well defined and that it "is a bi-product of a well-implemented intervention and is made evident through data collected via fidelity measures" (p. 1140). Therefore, because the core components of the program central to my study have been

well established, I restrict my examination to adherence, exposure, quality, and participant responsiveness.

Representation of the Components of FOI

Research that has examined the components of FOI has largely focused on adherence and exposure, with quality and participant responsiveness receiving far less attention (Dane & Schneider, 1998; Dusenbury et al., 2003). For example, Dane and Schneider's (1998) review found that of the studies that documented the degree to which program procedures were implemented, adherence and/or exposure were found in nearly every one, quality was found in only about a quarter of them, and participant responsiveness was represented in less than a tenth of them. Mihalic (2004) and Durlak and Dupre (2008) found nearly the same degree of representation among the FOI components—that is, adherence and exposure were much more common, with quality and participant responsiveness being represented in only a fraction of the studies surveyed. More recently, Gould, Dariotis, Greenberg, and Mendelson (2016) found that of the FOI studies they reviewed, exposure was the most common (48%), followed by adherence (20%), participant responsiveness (16%), and quality (16%). Although Dane and Schneider (1998) recommended that all dimensions be measured in order to get a complete picture of the program and FOI, Century et al. (2010) suggested that the selection of each dimension is determined by the needs and requirements of the individual study. Also, Durlak and DuPre (2008) noted that due to the complexities and time commitments of measuring implementation, most researchers focus only on a few variables that they deem as most critical. Azano et al. (2011) also noted that assessing all components is not necessary and that "researchers should attend to those components that are of interest to their study" (p. 696).

Relationship Among the Components of FOI

Carroll et al. (2007) hypothesized that there may be interactions among the components of FOI in predicting study outcomes and that research is needed to validate the connectedness of the components. Resnicow et al. (1998) pointed out that research has not examined concurrent or discriminant validity for the different measures of the components of fidelity, much less examined whether one component is a better predictor of outcomes than another, making it difficult to determine if one or another component is more valid. For example, the quality of delivery may affect participant responsiveness to program activities, or participants' adherence to program components may increase the quality in which they implement the components of the program. O'Donnell (2008) posited that the relationship between the dimensions of FOI is complex and often difficult to operationalize and conceptualize; therefore, a systematic review of the factors and strategies that optimize implementation, and more rigorous research and evaluation of the topic, is needed. Finally, Carroll et al. (2007) proposed that the inclusion of an explanation of the interrelationship between the components would add valuable information to the FOI literature. Hence, my current study offers a timely addition to the literature by providing more insight about the relationship among the FOI components.

Measuring FOI

There are several methods and combination of methods that can be used for measuring FOI; these can be direct or indirect. Direct methods, typically observations, examine the critical components (i.e., the elements of a program) that are considered essential to the outcomes of the program. Indirect methods can include self-reports, interviews, or the examination of artifacts. Typically, adherence and exposure are most often measured using indirect methods, such as teacher checklists indicating which topics were covered (adherence) and training attendance records (exposure)—perhaps one reason that they are much more represented in the FOI

literature. Studies have examined implementation quality with indirect methods, such as using teacher self-reports (e.g., see Hallfors & Godette, 2002; Ringwalt et al., 2002); however, Hansen and McNeal (1999) suggest that observations are a preferred method because of teacher self-reporting bias. Ross, McDougall, Hogaboam-Gray, and LeSage (2003) also offered that self-reports cannot accurately measure some aspects of practice, such as probing deep conceptual understanding, or that teachers may not have the same understanding as the developers of the items included on a self-report measure. Furthermore, when using observations to measure implementation quality, Sanchez et al. (2007) concluded that to produce a good measure of quality, observers should be skilled in the intervention or extensively trained in the observation protocol.

There are disadvantages to observations. First, they require extensive resources and time. Second, there are some dimensions of program components that may be difficult to observe, such as participant understanding (Ruiz-Primo, 2006). Finally, observations can have a reactivity measurement effect on the results, which may be due to the fact that the observant may take more time to prepare for the particular lesson or activity that is being observed than they might otherwise.

Self-report measures, such as checklists or questionnaires, can be used to determine the extent to which a program was implemented with adherence, the extent to which participants were responsive to the intervention, or the amount of time spent using the program activities. Self-reporting can have an advantage in that it provides a relatively accurate picture of classroom practice (Ross et al., 2003), as well as being easier to administer and analyze. Carroll et al. (2007) proposed that self-report measures can help evaluate the extent to which teachers or other participants are buying in to the responsibilities and requirements that are needed for the

implementation to be of value and to affect results. However, self-report measures may miss the physical or social settings in which something is being implemented, which in turn may affect the extent to which the results are favorable or consistent with actual implementation. This becomes critical when you examine the correlations between self-report measures and actual practice (Ross et al., 2003). An obvious limitation to self-report measures is the likelihood of potential discrepancies between what people say they do and what they actually do, which may be more of a measurement of social desirability than of actual performance.

Validity

According to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 2014), "Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests"(p. 11). Validity is, therefore, the most fundamental consideration in developing and evaluating tests. Cronbach (1971) described validation as the process by which a test developer or test user collects evidence to support the types of inferences that are to be drawn from test scores.

Over the last decade many researchers have identified critical steps in FOI development and measurement (Bond, et al., 2000; Century et al., 2010; Mowbray et al., 2003; O'Donnell, 2008), but validation and the methods for validating fidelity measures are still unclear. According to Mowbray et al. (2003) there are five different approaches that have been used to assess reliability and validity in the literature on FOI. Reliability has been assessed across respondents, calculating the extent of inter-rater agreement through coefficient kappa, intra-class correlations (ICC), percent agreement, or Pearson correlations (Clark & Watson, 1995; Henggeler, Schoenwald, Liao, Letourneau, & Edwards, 2002; Weisman et al., 2002). Reliability has also been assessed

using measures of internal consistency (e.g., Cronbach's alpha). The second approach, which focuses more on validity, has involved examining the internal structure of the data empirically and in relationship to expected results such as through exploratory and confirmatory factor analysis (Henggeler et al., 2002), or cluster analysis (Mills & Ragan, 2000). The third approach is the method of *Known Groups*, in which one examines differences in FOI scores across programs that are expected to be different (Bond et al., 2000; Hernandez et al., 2001; Lucca, 2000; Teague, Bond, & Drake, 2003). Typically, this involves a comparison of the new intervention compared to a traditional intervention. Convergent validity is the fourth approach that has been used in FOI validation. In the convergent validity approach, the focus is on examining the extent of agreement between two different sources of information about the program and its operations. For example, Blakely et al. (1987) compared records and documents with on-site observations and Macias, Propst, Rodican, and Boyd (2001) examined self-ratings of compliance with clubhouse standards on the Clubhouse Research and Evaluation Screening Survey (CRESS) to the results from on-site certification procedures, comparing CRESS scores of certified and non-certified agencies (Mowbray et al., 2003). The final approach, examining the relationship between fidelity measures and participant outcomes, is probably the most commonly used validation approach in research on interventions. When FOI components function in mediating a program delivery's effects on the outcome variables, there is some evidence for the validity of their use.

The current study is designed to examine the discriminant and convergent validity among the components of FOI across four data collection instruments. For this purpose, I use the multitrait-multimethod (MTMM) approach (Campbell & Fiske, 1959). The primary purpose of the approach as presented by Campbell and Fiske was to describe methods for collecting evidence

about the validity of inferences from data collected with instruments. The purpose of my use of the approach in this study, however, is not simply to provide validity evidence about instrumentation but rather to focus on the implications of my findings for the broad study of FOI. Due to the extensive instrument development process for the instruments that I use in my study, there is a high level of confidence of the psychometric quality of the instruments and that the inferences about the score interpretations on the FOI instruments are valid. However, due to the nature of the MTMM approach, each comparison within the matrix is a unit of both trait and method. Therefore, simply as a byproduct of focusing on the validity of the traits of interest, the findings are provided in light of the validity of the methods used to measure the traits.

Discriminant and Convergent Validity

A comprehensive explanation of discriminant and convergent validity is provided by Michael Kane (2006). Kane (2006) states that "*convergent validity evidence* (Campbell & Fiske, 1959) is evaluated in terms of the correlations between different measures of a trait. If these correlations are low, at least some of the measures are not adequately representing the target domain" (p. 36). Conversely, if the correlations between different measures designed to measure the same trait are high, then evidence of convergent validity is supported (Kane, 2006). Kane (2006) goes on to state that "empirical evidence on the relationship between distinct traits has been termed *discriminant validity evidence* (Cronbach, 1971)… If two measures are expected to be strongly related (e.g., because they are measures of the same trait), a strong empirical relationship between them provides evidence for the proposed interpretation. If two traits are expected to be unrelated, a strong empirical relationship counts against the proposed interpretation" (p. 39). Further, Kane (2000) states that "Campbell and Fiske (1959) proposed that the various kinds of correlational evidence relevant to the validation of a set of trait measures be considered within the context of a multitrait-multimethod matrix" (p. 39). In my study, including the different

components of FOI (i.e., adherence, quality, exposure, and participant responsiveness) and the different methods (e.g., observations and self-report measures) used to collected FOI data into one analysis will result in "a richer set of conclusions than could be derived from several separate analyses investigating specific concerns" (Kane, 2006, p. 39).

Improving Standards for Studying FOI

A review of the literature shows several gaps in our understanding about measuring and researching FOI. First, there has yet to evolve a universally acknowledged definition of FOI. Given how long that FOI has been of importance to designers (e.g., farmers of the 1920s and 30s), this lack of consensus may be due to the diversity of the different FOI fields of study (i.e., public health, judiciary, psychoanalysis, education, agriculture, and so forth), as well as the value that is placed on a set of specific FOI components that are important to each field. Second, of the studies that examine FOI, there is consistent underrepresentation of some of the FOI components compared to others. This limits the ability of researchers and evaluators to fully describe the strength that one component may have over another in predicting positive study outcomes. The unequal representation of the FOI components in much of the research may be the result of the logistical and financial burdens that many studies face in simply being able to examine their variables of interest, much less examine other variables that may help interpret the findings. Funding levels and logistics can affect the number of variables that any study can examine but improving our understanding about the components of FOI may increase the efficiency of FOI research by determining what particular components are central to a particular intervention's success. Further, as a result of the lack of consistency of FOI definitions, as well as the varied methods used to collect FOI data, there are no standards for the number of dimensions that need to be measured to get a complete picture of implementation, or any agreement on the best method for collecting FOI data. Therefore, by examining the convergent and discriminant

validity of the different components of FOI, as measured on different instruments, we will have a better understanding about the nature of the components and how best to measure them in a particular research setting.

CHAPTER 3

METHODS

The purpose of my study is to investigate the convergent and discriminant validity among four components of fidelity of Implementation (FOI)—adherence, exposure, quality, and participant responsiveness—for the purpose of deepening evaluators' and researchers understanding of FOI. The majority of the data that I used were collected as part of a larger study that examined the extent to which an inquiry-based science professional development (PD) positively affected science teacher instruction of an aquatic-centered science program (see Seraphin, 2014). I used data collected from three instruments designed for the project evaluation of the larger study. I also used data collected from an inquiry-based science observation protocol. I organized the data from all four instruments into correlation matrices to analyze the relationship among the four FOI components.

Research Design

I used a correlational research design to examine the extent to which the four FOI components were valid and distinct. Correlational research is used to investigate the relationships among variables, describe these relationships without attempting to influence the variables, and improve our understanding of the variables under examination (Fraenkel & Wallen, 2003). I used a multitrait-multimethod (MTMM) matrix (Campbell & Fiske, 1959) as a means to examine the correlations among the different FOI components as they were measured by the different instruments. I substitute the word *component* for the word *trait* in the remainder of my study. I continue use the acronym MTMM, however.

The key concept of the MTMM matrix is to determine the convergent and discriminant validity of multiple components as they are measured across multiple methods. The key evidence for convergent and discriminant validity is determined by the magnitude of the correlations

among the method-component scales. Convergent validity is shown when there is a strong relationship between a single component as it is measured across multiple methods (Kane, 2006; Marsh & Hocevar, 1984). Discriminant validity is shown when there is a moderate relationship among multiple components as they are measured on a single method and when there is a weak relationship among multiple components as they are measured across multiple methods (Kane, 2006; Marsh & Hocevar, 1984; Muis, Winne, & Jamieson-Noel, 2007). The logic of the MTMM matrix and the procedures for generating the MTMM matrix were an attempt by Campbell and Fiske to statistically develop the concept of a *nomological network*—a representation of the components of interest and the interrelationships among and between the components that provides the theoretical framework for studying a phenomenon—the foundation for construct validity (Cronbach and Meehl,1955).

MTMM Matrix

In Figure 3.1, I present an example of the MTMM matrix that shows the expected correlational levels among the method-component scales that would be observed if convergent and discriminant validity is supported (Salkind, 2010, p. 851). In referring to the example provided in Figure 3.1, I describe the different aspects of the matrix in the order of their expected correlation level (i.e., highest to lowest). First is the *reliability diagonal*, which I present in the figure with the term "very high" in parenthesis and highlighted in light grey. The reliability diagonal is used to present all of the *monocomponent-monomethod* comparisons. These provide the internal-consistency reliability estimates for each of the method-component scales. A method-component scale contains all of the items on a particular method that are intended to measure the component of interest. The internal-consistency reliability estimates are obtained by calculating Cronbach coefficient alpha. These values are expected to be very high (relative to the other correlations in the matrix) and provide the foundation for which to examine validity.

		Method 1			Method 2			Method 3		
	Component	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Method 1	A_1	(Very High)								
	B_1	Moderate	(Very High)							
	C_1	Moderate	Moderate	(Very High)						
Method 2	A_2	High	Low	Low	(Very High)					
	B_2	Low	High	Low	Moderate	(Very High)				
	C_2	Low	Low	High	Moderate	Moderate	(Very High)			
Method 3	A_3	High	Low	Low	High	Low	Low	(Very High)		
	B_3	Low	High	Low	Low	High	Low	Moderate	(Very High)	
	C_3	Low	Low	High	Low	Low	High	Moderate	Moderate	(Very High)

Figure 3.1. Example of a MTMM showing good convergent and discriminant validity (Salkind, 2010).

Second are the *validity diagonals*, which contain all of the *monocomponent-heteromethod* comparisons. I present these comparisons in the figure with the term "High" and highlighted with darker grey. These are the correlations between a single component as it is measured on multiple methods—the correlation values that provide the evidence for convergent validity of a component. Third are the *heterocomponent-monomethod triangles*, which I presented in the figure with the term "Moderate" and highlighted with slightly lighter grey than the validity diagonals. The three heterocomponent-monomethod triangles shown in Figure 3.1 are the correlations among multiple components as they are measured on a single method. These correlation values provide partial evidence for convergent validity of a component. The monocomponent-heteromethod correlations, along with the monocomponent-monomethod correlations make up the *monomethod blocks*. Fourth are the *heterocomponent-heteromethod triangles*, which I presented in the figure using the term "Low" with no highlighting. These are the values from the correlations among multiple components as they are measured across multiple methods. These correlation values also provide partial evidence for convergent validity

of a component. These heterocomponent-heteromethod correlations, along with the monomethod-heterocomponent correlations make up the *heteromethod blocks*.

Campbell and Fiske criteria. Campbell and Fiske (1959) established four criteria that need to be met to determine the convergent and discriminant validity of the components. The first criterion is used to determine convergent validity, and the last three criteria are used to determine discriminant validity. The first criterion is that the correlation value between a single component measured on multiple methods (monocomponent-heteromethod comparison) should be statistically significant from zero and large. The second criterion (the first criterion for discriminant validity) is that the correlation values among multiple components measured on a single method (heterocomponent-monomethod comparisons) should be greater than the correlation values among multiple components measured on multiple methods (heterocomponent-heteromethod comparisons) but less than the correlation values from the monocomponent-heteromethod comparisons (i.e., the comparisons used to determine convergent validity). The third criterion (the second criterion for discriminant validity) is that the correlation values among multiple components measured on multiple methods (heterocomponent-heteromethod comparisons) should be less than the correlation values among the multiple components measured on a single method (heterocomponent-monomethod comparisons). The fourth criterion (the third criterion for discriminant validity) is that the pattern of component intercorrelations should be similar among the sets of heterocomponent-monomethod comparisons and the sets of the heterocomponent-heteromethod comparisons. The expected correlation levels that I presented in Figure 3.1 (Salkind, 2010) are based on the hierarchical pattern of the Campbell and Fiske (1959) criteria.

MTMM analysis. To populate the MTMM matrix, I calculated the correlations between the method-component scales using the Pearson product-moment correlation coefficient (*r*). I chose to use the Pearson correlations based on the assumption of a linear relationship between the components, the small sample size from which my data were collected, and to specifically examine the individual method-component scale comparisons in light of the Campbell and Fiske (1959) criteria.

Researchers have used other procedures for analyzing the MTMM, such as *analysis of variance* (ANOVA) (King, Hunter, & Schmidt, 1980) and *confirmatory factor analysis* (CFA) (Werts & Linn, 1970). These procedures were developed to address some of the limitations of the Campbell and Fiske criteria, such as providing a means for separating the component and method variances (Schmitt & Stults, 1986). However, there are several restrictions to using these procedures that made them unsuitable in my study. First, I did not have a sufficiently large sample size needed for these procedures. Second, and more importantly, these procedures provide only general estimates of the component and method variances and do not provide a means for evaluating the individual method-component- scales—the focus of my study.

Context

The data used in my study were collected during the third and final year of the larger Teaching Science as Inquiry-Aquatic (TSI-A) PD project. The data were collected from all of the Year 3 teacher participants who were located on the islands of Oʻahu and Kauaʻi. The PD involved a series of four modules that consisted of in-person trainings and online learning support; it was developed to help teachers become successful facilitators of scientific inquiry within the context of aquatic science. The PD modules were based on the TSI pedagogical framework developed at the University of Hawaiʻi at Mānoa's Curriculum Research & Development Group (CRDG).

The modules were organized by four primary themes spaced throughout the school year: physical, biological, chemical, and ecological aquatic science. Each of the modules consisted of a two-day workshop, an in-person follow-up training, and a group online follow-up. The PD targeted teachers of heterogeneous groups of students in elementary, middle, and high schools throughout the state of Hawai‘i. The primary goals of the project were to increase teachers' content knowledge in aquatic science, improve teachers' science process and pedagogical knowledge, and improve student content knowledge and understanding of the nature of science.

Participants

The participants in this study included the 28 elementary, middle, and high school teachers who voluntarily participated in Year 3 of the study. There were 13 teachers from Kaua‘i and 15 teachers from O‘ahu. In Table 3.1, I present the number of teachers from each island and grade band level, and show the range in the number of years they have been teaching science.

Project Evaluation

The evaluation of the TSI-A project was a mixed-method study that collected data for both formative and summative evaluation purposes. The questions guiding the evaluation of the TSI-A project focused on project training, project implementation, and the effects that the project had

Table 3.1
Number of Participating Teachers and the Number of Years They Have Been Teaching Science , by Island and Grade Band

			No. of years teaching science			
Island	Grade band	No of teachers	1-5	6-10	11-15	≥16
O‘ahu	Elementary	2		1	1	
	Middle	8	2	3	1	2
	High	5	2	1	2	
Kaua‘i	Elementary	1	1			
	Middle	6	3	2		1
	High	6	2	1	2	1
	Total	28	10	8	6	4

on teacher and student learning. The four-member evaluation team, which included the evaluation PI, evaluation project manager, and two graduate assistants, concentrated on instrument development and pilot-testing and provided formative-evaluation feedback to project developers during the first two years of the project. The third and final year was dedicated to collecting data for the study's summative evaluation component (see Seraphin, 2014).

Data Sources

The list of instruments used for my study and the FOI components they measured are presented in Table 3.2. I provide an in-depth description of each of the instruments in the subsequent sections. Three of the instruments described in my study (Teacher Activity Reflection, Post-Cohort Questionnaire, and Teacher Interview) were used to collect data for the TSI-A project evaluation. The Inquiring into Science Instruction Observation Protocol (ISIOP) (Minner & DeLisi, 2012) was chosen specifically for my study.

Fidelity of Implementation Item Identification

The FOI items that I selected for analysis was accomplished through a series of steps intended to systematically match each of the items to one of each of the four FOI components, if appropriate. This process ensured that only the FOI components items specifically related to teacher implementation of the TSI-A project components were included. In the following section, I discuss the item selection process.

Table 3.2
List of Instruments and Components They Were Intended to Measure

	FOI component			
Instrument	Adherence	Exposure	Quality	Participant responsiveness
Teacher Activity Reflection	X		X	X
Teacher Interview				X
Post-Cohort Questionnaire	X	X		X
Inquiring into Science Instruction Observation Protocol		X	X	

Fidelity of Implementation Scale Selection Steps

The selection of items that I used in my study occurred in three steps. First, I included all the items from the four instruments into a MS Excel spreadsheet and distributed the spreadsheet to each of the four members of the TSI evaluation team. Each member independently, and on his or her own schedule, assigned items to one of the four FOI components based on the component definitions that I presented in Chapter 2. Second, after each member completed his or her assignment of the instrument items to one of the four FOI components, we convened as a group to discuss and finalize the assignments. For this meeting, I combined each member's item assignment into separate columns on a single spreadsheet and displayed it on a projector for group review. Third, as a group, we reviewed each FOI component assignment item-by-item. When there was a lack of complete agreement between any of the members, we reviewed the definitions, and the members discussed the extent to which an item best represented an FOI component until all differences were reconciled. A total of 141 items were finally selected for inclusion in my study. In Table 3.3, I present the number of each of the FOI component items represented on each of the four instruments.

Evaluation Instruments Examined for this Study

In this section, I provide a detailed description of each of the instruments that were used to collect FOI data for my study. I also discuss how the data were aggregated to create the final FOI component-method scales for use in the MTMM matrix.

Target Activity Reflection

The Target Activity Reflection (hereafter *Reflection*) is a self-report measure that was designed to gather information about the teachers' perceptions of their implementation of the 15 target activities they were taught over the course of the school year. The Reflection was

Table 3.3
Number of Items Addressing Each of the FOI Components by Instrument

Instrument	No. of items by FOI component			
	Adherence	Exposure	Quality	Participant responsiveness
Teacher Activity Reflection	8		17	9
Teacher Interview				5
Post-Cohort Questionnaire	4	10		59
Inquiring into Science Instruction Observation Protocol		5	24	
Total no. of items	12	15	41	73

developed over multiple iterations with input from both the evaluation and project development team members during the first two years of the project. As the final year progressed, there were slight variations in the Reflection to mirror the teachers' growing knowledge and skill set about the project's components and topics. For example, the Reflections for Modules 3 and 4 included additional items to gather more focused information about the teachers' use of the *phases* and *modes* of inquiry—primary components of the TSI program model. This was because the teachers were still learning about the phases and modes in Modules 1 and 2, and they were not expected to have the level of knowledge needed for full implementation compared to Modules 3 and 4.

The phases of inquiry are central aspects of the general TSI program (Seraphin, 2014). The five phases are connected but are not intended to be sequential; the emphasis is on the possibility of multiple logical progressions rather than having specific procedural steps in the inquiry process. The TSI-A final report (Seraphin, 2014) provides an example of how this non-sequential interaction between the phases may occur, "initiation can occur at the beginning of a lesson, but it can also occur throughout the course of investigation as students re-initiate by experiencing anomalies, asking questions, or considering new information. An encountered difficulty interpretation can redirect the learning cycle, leading to the need for invention of new processes

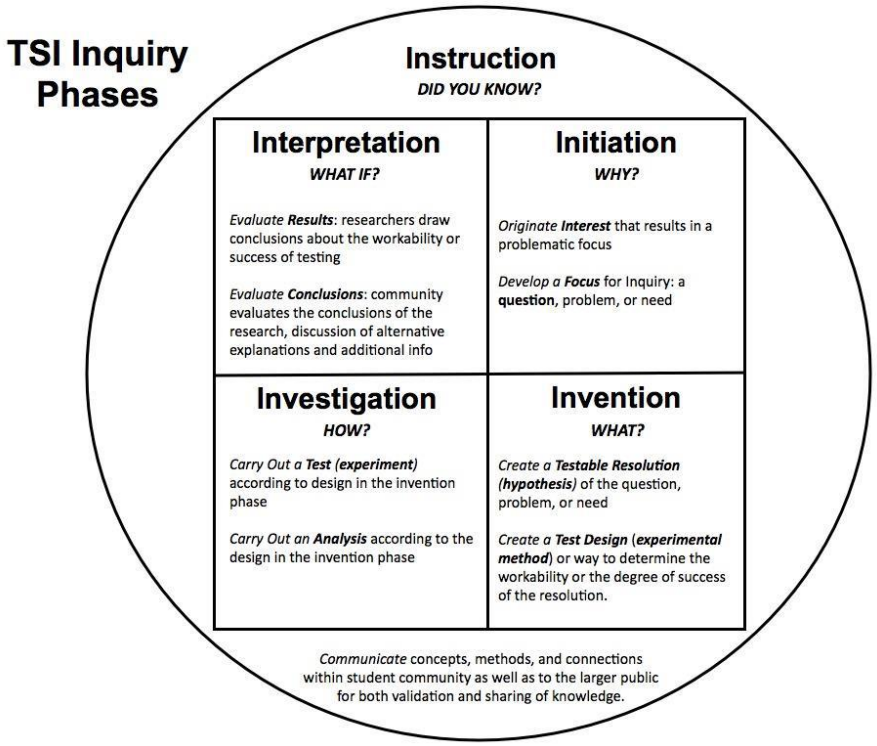

Figure 3.2. The TSI square-in-circle phase diagram (from Seraphin, 2014).

or ideas to be investigated" (p. 9). In Figure 3.2, I present the square-in-circle diagram taken from the TSI final report, which is intended to reflect what happens in realistic scientific processes. The modes of inquiry are another central aspect of the TSI program and are intended to reflect the different ways in which to do scientific inquiry. In Table 3.4, I present the description, taken from the TSI final report, of each of the ten modes that are addressed in the TSI inquiry model.

There were four Reflections administered for the activities that were covered in Module 1, four for Module 2, four for Module 3, and three for Module 4. The teachers were required to complete at least three of the four Reflections for each of the *target activities* covered in Modules 1 through 3; teachers completed all three Reflections for the three activities covered in Module 4. In Table 3.5, I present the response rate for the 15 Reflections. The Reflection was completed by

Table 3.4
The Modes of Inquiry Addressed in TSI-A (from Seraphin, 2014)

Mode (Inquiry learning through use of ____)	Description Search for new knowledge…
Curiosity	in external environments through informal or spontaneous probes into the unknown or predictable
Description	through creation of accurate and adequate representation of things or events
Authoritative knowledge	through discovery and evaluation of established knowledge via artifacts or expert testimony
Experimentation	through testing predictions derived from hypotheses
Product Evaluation	about the capacity of products of technology to meet valuing criteria
Technology	in satisfaction of a need through componention, production and testing of artifacts, systems, and techniques
Replication	by validating inquiry through duplication; testing the repeatability of something seen or described
Induction	in data patterns and generalizable relationships in data association—a hypothesis finding process
Deduction	in logical synthesis of ideas and evidence—a hypothesis making process
Transitive knowledge	in one field by applying knowledge from another field in a novel way

Table 3.5
Response Rates for the 15 Activity Reflections

Module	Activity Reflection	*N*
1	Practices of Scientists	28
	Soda and Scientific Reasoning	23
	Density Bags	28
	Kinesthetic Moon Model	5
2	Electrolysis	15
	Conductivity	13
	Water Properties	28
	Phases and Modes of Scientific Practice	28
3	Modeling Microevolution	22
	Scientific Language	28
	Fish Form and Function	16
	Phases and Modes of Scientific Practice	18
4	Own Lesson	28
	Sampling for Abundance	28
	Sampling Design	28

the teachers, online via SurveyMonkey, immediately after they had completed their implementation of a target activity. The Reflection items consisted of Likert scales, open-ended, and multiple choice response options. The Reflection items were separated into sections that covered *Activity Implementation, Use of Inquiry, Effect of Professional Development,* a section that focused on teachers' use of the *phases and modes* of inquiry, and a section for final comments. The Activity Implementation section included items that asked the teachers to rate the extent to which they followed TSI-A activity procedures, covered aquatic centered topics and questions, and the perceived effects that the activity content had on student learning. The Use of Inquiry section focused on teachers' self-ratings of how well they covered the topics; the extent that they guided their students through the phases and modes of inquiry; and the effect that using inquiry had on their understanding of inquiry. The Effect of Professional Development section asked the teachers to rate the extent to which the PD covered the content and to retrospectively rate the extent to which the PD affected their knowledge and confidence after they had completed teaching the activity. The section about phases and modes asked the teachers to indicate the

extent of their use of the phases and modes, as well as the extent to which the phases and modes affected student learning. A full description of the PCQ is provided in the TSI-A final report (Seraphin, 2012).

The Reflection addressed three of the four FOI components of adherence, quality, and participant responsiveness. To prepare the items for the MTMM matrix, I first calculated the mean response on each individual item across the Reflections within each Module. I illustrate this process using the following formula: $Item_{k,t,m} = (Item_{k,t,r1} + \ldots Item_{k,t,rn})/n$, where *k* represents the Item, *t* represents the teacher, *m* represents the module, and *r* represents the

Reflection with *n* reflections in which Item *k* occurred within module *m*. So, $Item_{k,t,m}$ is Teacher *t*'s mean response on Item *k* across all reflections in which it was present in Module *m*. This first step was done to account for the differential response rates for each of the Reflections (i.e., teachers were only required to complete three of the four reflections in Modules 1 through 3; thus, some Reflections had a significant number of missing values); this also ensured that each item had a value that could be used in the correlation analysis. Next, I calculated the teachers' mean responses to that item across the four modules. I illustrate this using the following formula: $Item_{k,t} = (Item_{k,t,m1} + Item_{k,t,m2} + Item_{k,t,m3} + Item_{k,t,m4})/4$. $Item_{k,t}$ is Teacher *t*'s mean response on Item *k* across all modules. Finally, I calculated the teachers' mean responses across all the module-level items' means to get each teacher's final composite score for each of the FOI components. I illustrate this using the following formula: $Item_s = (Item_{k,t1} + \ldots Item_{k,tn})/n$, where *s* represents the Item's scale score. Once the composite scores were created for each teacher for each item, I grouped the Item *k* means by the FOI component they were intended to measure. Next, I calculated Cronbach's alpha coefficient among the other item composite scores representing the same FOI component to determine the scale's internal-consistency reliability.

To ensure that the group of items that represented one of the three components addressed on the Reflection had adequate internal consistency, I set a cut value of $\alpha \geq .70$ (Nunnally & Berstein, 1994). If the alpha value was less than .70, I examined each of the individual item's correlation with the total score, identified the item with the lowest correlation with total, removed the item from the analysis, and recalculated the alpha coefficient. In addition to ensuring that the alpha coefficient was greater than or equal to .70, I also examined each of the items' correlation with the total of the scale. If an item's correlation with the total was less than .50 (Clark & Watson, 1995; Francis & White, 2002), the item was removed and Cronbach's

alpha was recalculated. If the alpha increased, and all other item correlations with the total remained greater than .50, the item was omitted from the final FOI component scale. This process was completed until the highest level of internal consistency was achieved for each of the FOI component scales. In Appendix A, Tables A.1 to A.3, I present the Reflection items that I used in my study; show which items from this list were selected for omission from the final scale; and provide the final internal-consistency reliability coefficients for each of the FOI component scales.

Post-Cohort Questionnaire

The Post-Cohort Questionnaire (PCQ) is a self-report measure that was administered to the teachers, online via SurveyMonkey, at the end of the school year after all the required activities had been implemented. The evaluation team designed the PCQ to get a general understanding of the teachers' experiences with the project. After several iterations over the course of the three-year TSI-A project, the final version included 100 6-point Likert-scale items that resulted in 14 scales. The evaluation team organized the final version into sections representing the 14 scales, including: (a) the value and relevance of the TSI-A PD, (b) the effect of the TSI-A PD on teaching, (c) the level of comfort of implementing the TSI-A activities, (d) the perceived effect that the TSI activities had on students, (e) the extent to which all steps were implemented, (f) how often the modes of TSI-A were included in the instruction, (g) the value of the modes in instruction, (h) comfort with using the modes after completing the TSI-A PD, (i) comfort with using the modes before completing the TSI-A PD, (j) how useful the TSI components were in teaching science as inquiry, (k) how often the TSI-A modes will be used in the future, (l) how often the TSI phases will be used in the future, m) familiarity with Ocean Literacy Principles, and n) the extent that the PD requirements improved the teachers understanding of inquiry. A full description of the PCQ is provided in the TSI-A final report (Seraphin, 2012). I analyzed only

the items from the PCQ that were identified as addressing one of the FOI components. A total of 52 items were selected from the PCQ for my analysis. The FOI components addressed in the PCQ are adherence, exposure, and participant responsiveness. The final scale scores for each of the FOI components were created by calculating the teacher means across each of the items that represented one of the three FOI components. As opposed to the Reflection, which required several steps of mean calculations to obtain final teacher composite scores, the PCQ scales were calculated by simply averaging across the specified items that represented one of the three FOI components.

Next, as with the procedure used in the Reflection, I calculated the internal consistency (Cronbach's coefficient alpha) of the items that were intended to comprise the FOI component scale. If the internal consistency of all the items included for a specific FOI component was found to be less than $\alpha = .70$, I examined each of the items' correlation with the total score, identified the item with the lowest correlation with total, removed the item from the analysis, and recalculated the alpha coefficient. In addition, I also examined each of the item's correlation with the total score. If an item's correlation with total was less than .50, the item was removed and Cronbach's alpha was recalculated. If the alpha increased, and all other item correlations with the total remained greater than .50, the item was omitted from the final FOI component scale. In Appendix A, Tables A.4 through A.6, I present the PCQ items that I used in my study; show which items from this list were selected for omission from the final scale; and provide the final internal-consistency reliability coefficients for each of the FOI component scales.

Teacher Interview

The purpose of the Teacher Interview (hereafter *Interview*) was to gather both qualitative and quantitative information about the teachers' experience with the different topics addressed in the TSI-A project. The interviews were conducted by me and another evaluation team member using

an interview script developed by the evaluation team. The interviews were conducted live (online via Blackboard) with 22 of the 28 teachers at the end of the school year. For the remaining six teachers who were unable to schedule live interviews, an identical written version was provided. The interview was conducted by reminding teachers about the major topics, asking them to provide ratings, on a *1* to *10* Likert scale about aspects of their implementation for each of the major topics, and then asked several open-ended questions that allowed them to expand on their ratings. After the interviews were completed, we independently performed a content analysis, based on an agreed protocol, and met to finalize the accuracy of our results and reconcile any differences. A full description of the Interview is provided in the TSI-A final report (Seraphin, 2012). Only the quantitative data provided by the interview were used to create the Interview component scale; however, I do reference some of the qualitative results to help in some of my interpretations of the MTMM correlation results. The final group of items selected from the Interview addressed only the FOI component of participant responsiveness. Similar to the PCQ, I created a final participant responsiveness scale score by averaging the means for each of the five items. I calculated the internal consistency to ensure the overall scale had a value of $\alpha \geq .70$. No items were selected for omission from this scale. In Appendix A, Table A.7, I present each of the items correlations with the total and provide the final internal-consistency reliability coefficient for the Interview participant responsiveness scale.

Inquiring into Science Instruction Observation Protocol

The purpose of the ISIOP was to collect FOI data by observing the teachers' implementation of a TSI-A target activity. Due to time and financial constraints, only the 15 teachers located on O'ahu were observed. I selected the ISIOP for conducting the observations due to the high level of research and validation that was done during its development and the extent to which it

measured inquiry-based science practices. The following summarizes the ISIOP as described by Minner and DeLisi (2012) in their user's manual.

> The purpose of the ISIOP is to assist evaluators and researchers in determining the extent to which quality pedagogical practices, including instruction that integrates scientific practices, are present in science classrooms…. The ISIOP reflects a comprehensive view of inquiry-oriented classroom practice, focusing on teaching indicators that have been either theorized or demonstrated to be associated with student learning—specifically those instructional practices that are exhibited during a given lesson…. The development and framework of the ISIOP is the result of an extensive review of the literature on inquiry-based instruction (e.g., Minner, Levy, & Century, 2010) and an examination of existing instruments…. This conceptual grounding in the literature provided the groundwork for the development of items contained in the protocol and established one early line of evidence for content validity of the items. (pp. 1–3)

One of the unique aspects of the ISIOP is the data-collection structure that distinguishes between what a teacher says in her interactions with students, the kinds of activities in which students engage, and the presence of scientific practices. The protocol reflects a *componentivist approach* to teaching, which focuses on engaging students physically and cognitively in the act of learning—a key component to inquiry-based instruction—rather than relying exclusively on the traditional teacher-directed approach.

I used the ISIOP to observe the teachers implementation of the *Sampling Design* target activity. This was one of the final three target activities taught during Module 4—the final module. I chose this particular target activity because it was designed to comprehensively address the different aspects of the TSI-A pedagogy (i.e., phases, modes, etc.) and was

implemented at a point in time when the teachers had completed all their training and were expected to be fluent in the TSI-A pedagogy.

ISIOP quality assurance. To ensure that there was an acceptable level of reliability of the data collected with the ISIOP, another TSI team evaluator participated with me in the training and the in-class observations. We became well-versed in the ISIOP User's Manual and participated in the online training provided by the ISIOP developers (Minner & DeLisi, 2012). The training consisted of several coding steps that were taken from actual videos of inquiry-based science classroom lessons collected during the ISIOP development period. The trainee is expected to check interrater reliability for (a) lesson events, (b) verbal practices and investigation experiences, and (c) the entire observation protocol that are covered on the ISIOP. The goal of the training was to ensure that the trainee had 70% interrater reliability (Kendall's τ) for each component of the ISIOP prior to implementing the ISIOP in a classroom setting. It was critical to have at least two trained observers because there was overlap between when the teachers scheduled their implementation of the target activity. That is, some teachers scheduled their implementation at about the same time on the same day. Due to this overlap in scheduling, we could not observe some of the teachers simultaneously. However, because we were both trained in the procedure, both of us could independently observe the teachers and still maintain a high level of reliability. In addition to participating in the training together, we simultaneously observed eight partial and full lessons together throughout the implementation period. We did this to ensure that we maintained high levels of reliability throughout the process. For the eight partial and full lessons that we observed together, we had a high level of interrater reliability (Kendall's $\tau = .85$). This provided evidence that we were adequately trained to observe the lessons. In most cases, the teachers' implementation of target activities took more than a single

class period, resulting in observations by only a single observer when both observers could not be present. In Table 3.6, I present the total amount of time (in minutes) that each teacher took to implement the activity. In this table, I also present the number of days that each teacher took to complete their implementation of the target activity, and use check marks to show which of the days were observed individually by me (Observer 1), which days were observed individually by the other evaluator (Observer 2), and which days we observed together.

The ISIOP data were coded using tally marks to indicate frequency of verbal practices; dichotomous (yes/no) options to indicate if an inquiry-based science topic was addressed; and Likert scales to indicate the extent to which a topic was addressed, as well as the extent to which the observant exhibited characteristics indicative of inquiry-based science instruction. From the available items, only items that addressed the FOI components of exposure and quality were identified as appropriate for my study.

Table 3.6
Duration of the Teachers' Implementation of the Target Activity

Teacher	Time (minutes)	Observer 1			Observer 2		
		Day 1	Day 2	Day 3	Day 1	Day 2	Day 3
1	119	✓	✓		✓		
2	87	✓			✓	✓	
3	93	✓					
4	120	✓	✓				
5	35	✓					
6	130	✓	✓		✓		
7	100	✓			✓	✓	
8	51	✓	✓				
9	85	✓	✓		✓		
10	83	✓	✓				
11	161	✓	✓	✓	✓	✓	✓
12	75	✓	✓				
13	143	✓	✓	✓		✓	✓
14	153				✓		
15	88	✓	✓				

As with the other instrument items, I created the final FOI component scales by calculating the teacher means across each of the items for each of the two FOI components. I then calculated the internal-consistency reliability coefficients, removed items that had low correlations to the total (less than .50) if the reliability was less than .70, and recalculated the reliability coefficients for the final component scales. The ISIOP quality scale included two items with item-total correlations less than .50 because removal of these items, in addition to the one item that was removed, would have resulted in only a single item scale. These two correlations with the total were very close to the .50 level (i.e., .48 and .49), which I deemed was acceptable. In Appendix A, Tables A.8 and A.9, I present the item list, the item that was selected for omission from the final scale, and the internal consistency of each FOI component scale.

FOI Method-Component Scale Analyses

The final FOI component scales are presented in two MTMM matrices. The first matrix included the scales for all 28 teachers across the three TSI-A evaluation instruments. The second matrix was developed because only a subset of the 28 teachers were observed on the ISIOP (i.e., the 15 Oʻahu teachers). It included the FOI component scales from the three TSI-A evaluation instruments, as well as the FOI component scales from the ISIOP. The first matrix included seven FOI component scales which resulted in a 7-by-7 matrix. These represent the FOI components addressed on the Reflection, the PCQ, and the Interview. These seven scales are the Reflection adherence, Reflection quality, and Reflection participant responsiveness scales; the PCQ adherence, PCQ exposure, and the PCQ participant responsiveness scales; and the Interview participant responsiveness scale. The second matrix included the seven scales from the first matrix and the two ISIOP scales: the ISIOP exposure and the ISIOP quality scales, resulting in a 2-by-9 matrix.

In Table 3.7, I present the 7-by-7 matrix and show the correlations each cell represents. For example, $r_{RA \cdot RQ}$ represents the correlation (r) between the Reflection (R) adherence (A) scale and the Reflection (R) quality (Q) scale. In this table, I have highlighted the different comparisons that will be used to determine convergent and discriminant validity using solid and broken lines.

The correlations that provide the evidence for convergent validity are enclosed in broken line rectangles. These are the monocomponent-heteromethod comparisons that are expected to be the highest correlations in the matrix and statistically significant from zero (Campbell and Fiske's criterion for convergent validity). As shown in Table 3.7, there are four monocomponent-heteromethod comparisons: one is used to show convergent validity for adherence (i.e., $r_{RA \cdot PA}$), and three are used to show convergent validity for participant responsiveness. The correlations that are used to address Campbell and Fiske's first criterion for discriminant validity are enclosed in solid line triangles. These are the heterocomponent-monomethod comparisons that are

Table 3.7
7-by-7 MTMM Matrix Showing the Correlations That Address the Campbell and Fiske (1959) Criteria for Convergent and Discriminant Validity

		Reflection			PCQ			Interview
		A	Q	PR	A	E	PR	PR
Reflection	A	$\alpha_{RA \cdot RA}$						
	Q	$r_{RA \cdot RQ}$	$\alpha_{RQ \cdot RQ}$					
	PR	$r_{RA \cdot RPR}$	$r_{RQ \cdot RPR}$	$\alpha_{RPR \cdot RPR}$				
PCQ	A	$r_{RA \cdot PA}$	$r_{RQ \cdot PA}$	$r_{RPR \cdot PA}$	$\alpha_{PA \cdot PA}$			
	E	$r_{RA \cdot PE}$	$r_{RQ \cdot PE}$	$r_{RPR \cdot PE}$	$r_{PA \cdot PE}$	$\alpha_{PE \cdot PE}$		
	PR	$r_{RA \cdot PPR}$	$r_{RQ \cdot PPR}$	$r_{RPR \cdot PPR}$	$r_{PA \cdot PPR}$	$r_{PE \cdot PPR}$	$\alpha_{PPR \cdot PPR}$	
Interview	PR	$r_{RA \cdot IPR}$	$r_{RQ \cdot IPR}$	$r_{RPR \cdot IPR}$	$r_{PA \cdot IPR}$	$r_{PE \cdot IPR}$	$r_{PPR \cdot IPR}$	$\alpha_{IPR \cdot IPR}$

A = adherence; E = exposure; Q = quality; PR = participant responsiveness
R = Reflection; P = PCQ; I = Interview

expected to have (a) moderate correlation values, (b) correlation values that are lower than the monocomponent-heteromethod correlation values, and (c) correlation values that are larger than the heterocomponent-heteromethod correlation values. There are a total of six monocomponent-heteromethod comparisons: three are the correlations among the components as measured on the Reflection, and three are the correlations among the components as measured on the PCQ. Finally, the correlations that are used to address Campbell and Fiske's second criterion for discriminant validity are enclosed in solid line rectangles. These are the heterocomponent-heteromethod comparisons that are expected to be the lowest correlation values in the matrix. As shown in Table 3.8, there are a total of 11 heterocomponent-heteromethod comparisons.

In Table 3.8, I present the 2-by-9 matrix and show what correlations each cell represents. I used the same solid and broken lines to highlight the different comparisons that I used in the 7-by-7 matrix (Table 3.7) to show how each cell is used to determine convergent and discriminant

Table 3.8
2-by-9 MTMM Matrix Showing the Correlations That Address the Campbell and Fiske (1959) Criteria for Convergent and Discriminant Validity

		ISIOP	
Instrument	FOI component	Exposure	Quality
ISIOP	Exposure	$\alpha_{ISE \cdot ISE}$	
	Quality	$r_{ISE \cdot ISQ}$	$\alpha_{ISQ \cdot ISQ}$
Reflection	Adherence	$r_{ISE \cdot RA}$	$r_{ISQ \cdot RA}$
	Quality	$r_{ISE \cdot RQ}$	$r_{ISQ \cdot RQ}$
	Participant Responsiveness	$r_{ISE \cdot RPR}$	$r_{ISQ \cdot RPR}$
PCQ	Adherence	$r_{ISE \cdot PA}$	$r_{ISQ \cdot PA}$
	Exposure	$r_{ISE \cdot PE}$	$r_{ISQ \cdot PE}$
	Participant Responsiveness	$r_{ISE \cdot PPR}$	$r_{ISQ \cdot PPR}$
Interview	Participant Responsiveness	$r_{ISE \cdot IPR}$	$r_{ISQ \cdot IPR}$

A = adherence; E = exposure; Q = quality; PR = participant responsiveness
R = Reflection; P = PCQ; I = Interview; IS = ISIOP

validity. As shown in Table 3.8, there are 2 heterocomponent-monomethod comparisons (broken line rectangles), 1 heteromethod-monocomponent comparison (solid line triangle), and 12 heterocomponent-heteromethod comparisons (solid line rectangles).

Descriptive Statistics for the FOI Component Scales

In Tables 3.9 and 3.10, I present the descriptive statistics for the FOI component scales used in each of my two MTMM matrices. To present the normality of the different scales visually, I also present the QQ (quantile-quantile) probability plots for each of the FOI component scales for all 28 teachers in Figures 3.3 to 3.9. In Figure 3.10 and 3.11 I present the QQ plots for the two ISIOP of exposure and quality only the 15 teachers who participated in the in-class observations. As shown in Tables 3.9 and 3.10, all FOI scales followed a relatively normal distribution (indicated by the Shapiro-Wilk $p > .05$) except for adherence as measured on the PCQ ($p = .01$) and exposure as measured on the ISIOP ($p = .02$).

Table 3.9
Descriptive Statistics for Final Composite Scores for All 28 Teachers Across Four Instruments

Instrument	FOI component	*M*	*SD*	*s.e.*$_M$	*Kurtosis*	*Skewness*	Shapiro-Wilk	
							W	*p*
Reflection[a]	Adherence	3.04	0.56	0.11	-0.15	0.24	.98	.82
	Quality	3.83	0.29	0.06	-0.39	0.33	.98	.84
	Participant Responsiveness	4.00	0.38	0.07	-0.50	-0.29	.96	.35
PCQ[b]	Adherence	5.04	0.64	0.12	-0.73	< -0.01	.90	.01
	Exposure	4.58	0.84	0.16	-0.75	-0.38	.94	.09
	Participant Responsiveness	5.20	0.53	0.10	-0.39	-0.47	.96	.31
Interview[c]	Participant Responsiveness	8.06	1.11	0.21	-0.50	-0.39	.95	.26

[a] 5-point Likert scale, where 1 = low rating and 5 = high rating; [b]6-point Likert scale, where 1 = low rating and 6 = high rating; [c]10-point Likert scale, where 1 = low rating and 10 = high rating

Table 3.10
Descriptive Statistics for Final Composite Scores for 15 Teachers Across One Instrument

Instrument	FOI component	*M*	*SD*	*s.e.*$_M$	*Kurtosis*	*Skewness*	Shapiro-Wilk	
							W	*p*
ISIOP[a]	Exposure	1.07	0.58	0.15	0.06	.96	.85	.02
	Quality	2.14	0.55	0.14	-0.99	-0.27	.95	.53

[a] 4-point Likert scale, where 0 = none and 3 = a lot

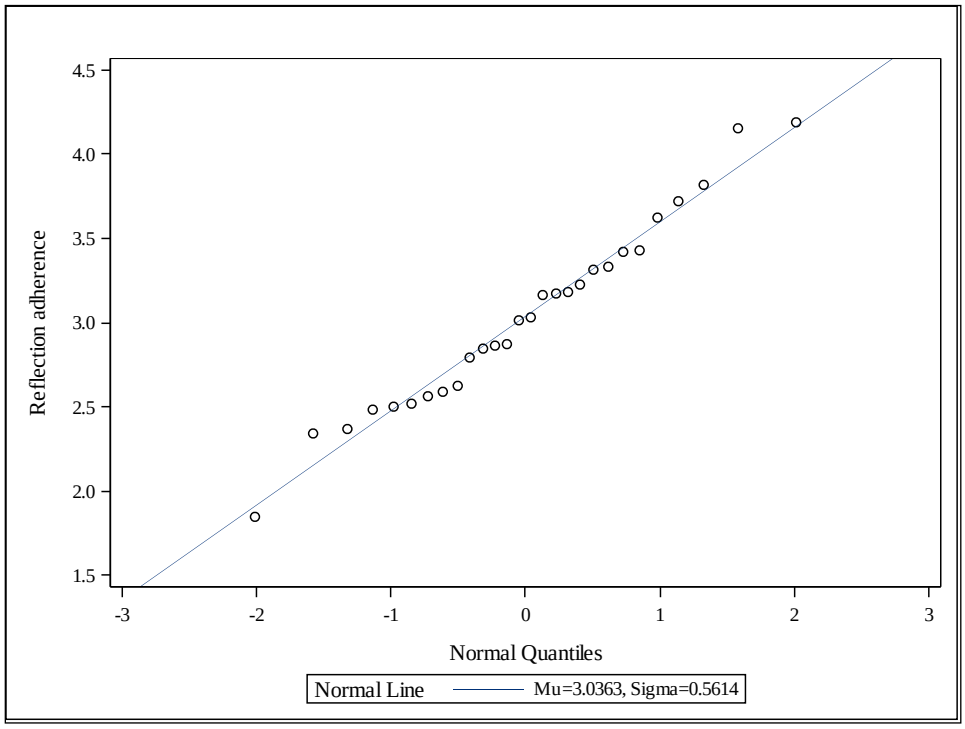

Figure 3.3. QQ plot for reflection adherence scale.

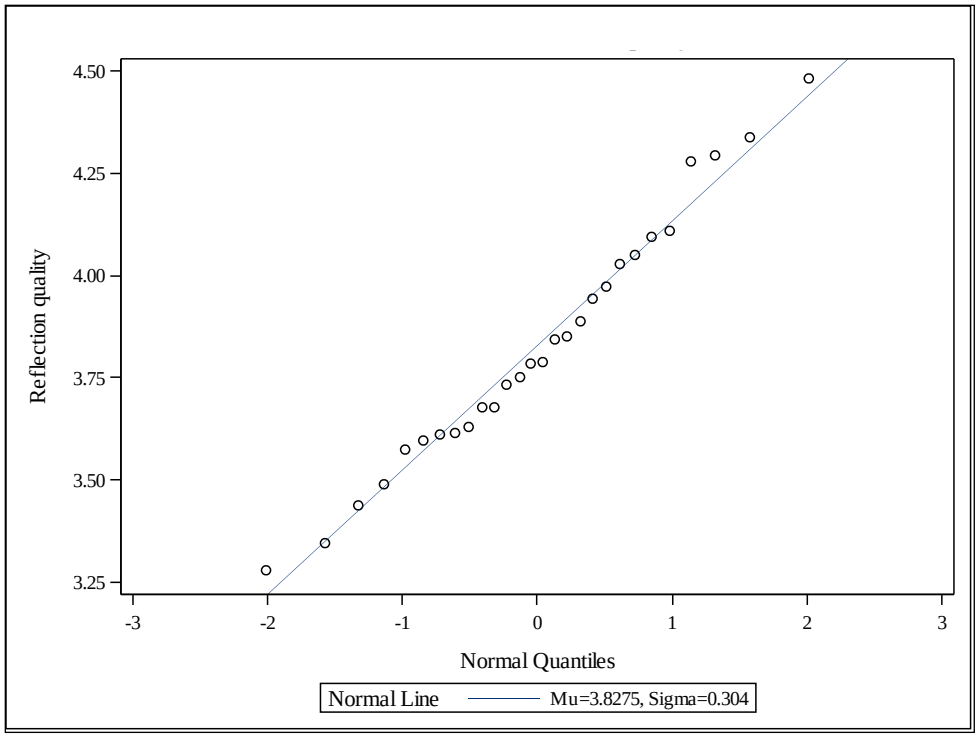

Figure 3.4. QQ plot for reflection quality scale.

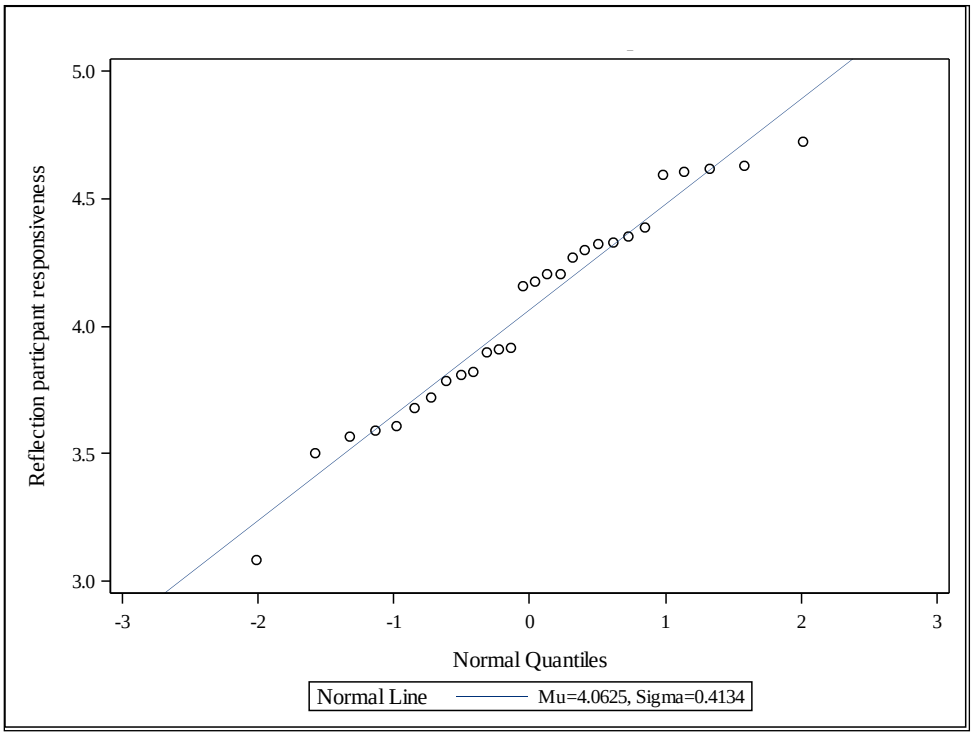

Figure 3.5. QQ plot for reflection participant responsiveness scale.

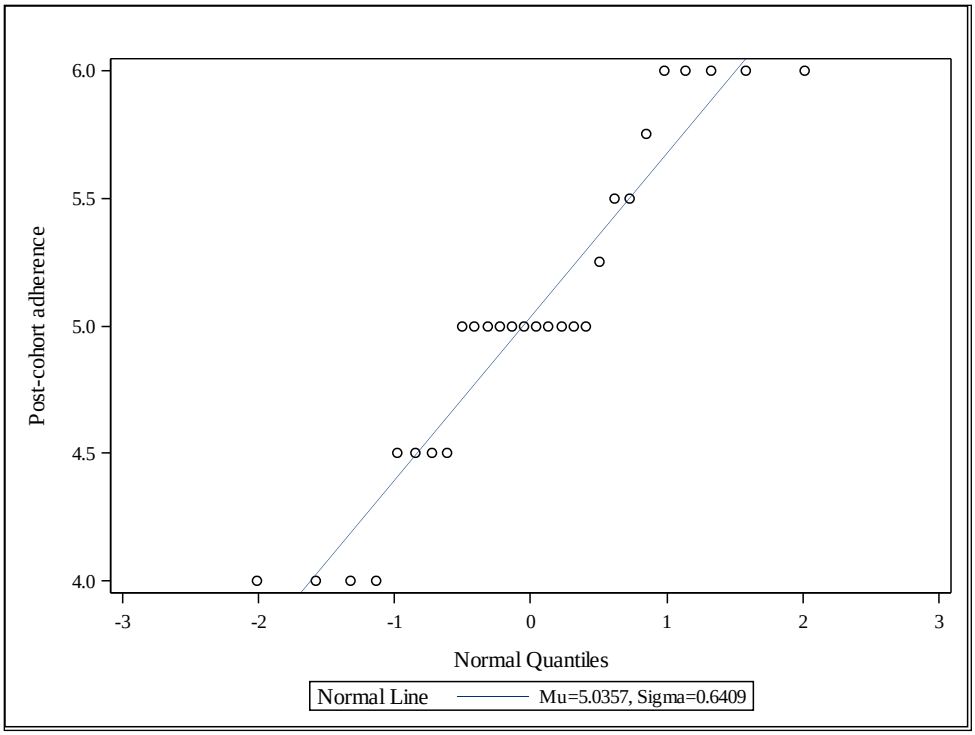

Figure 3.6. QQ plot for PCQ adherence scale.

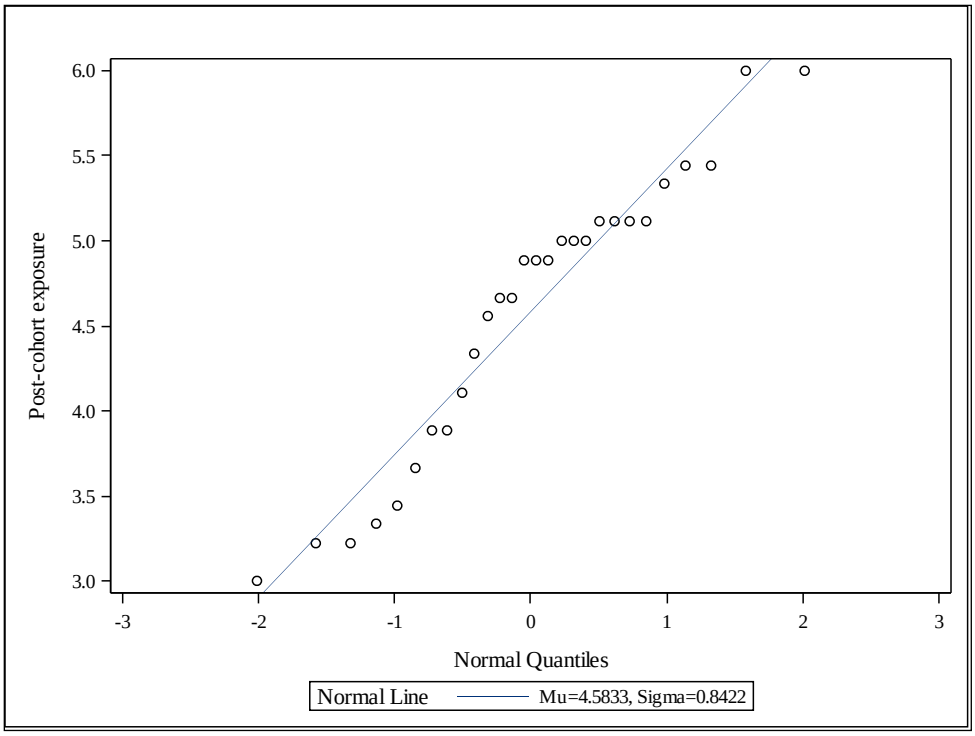

Figure 3.7. QQ plot for PCQ exposure scale.

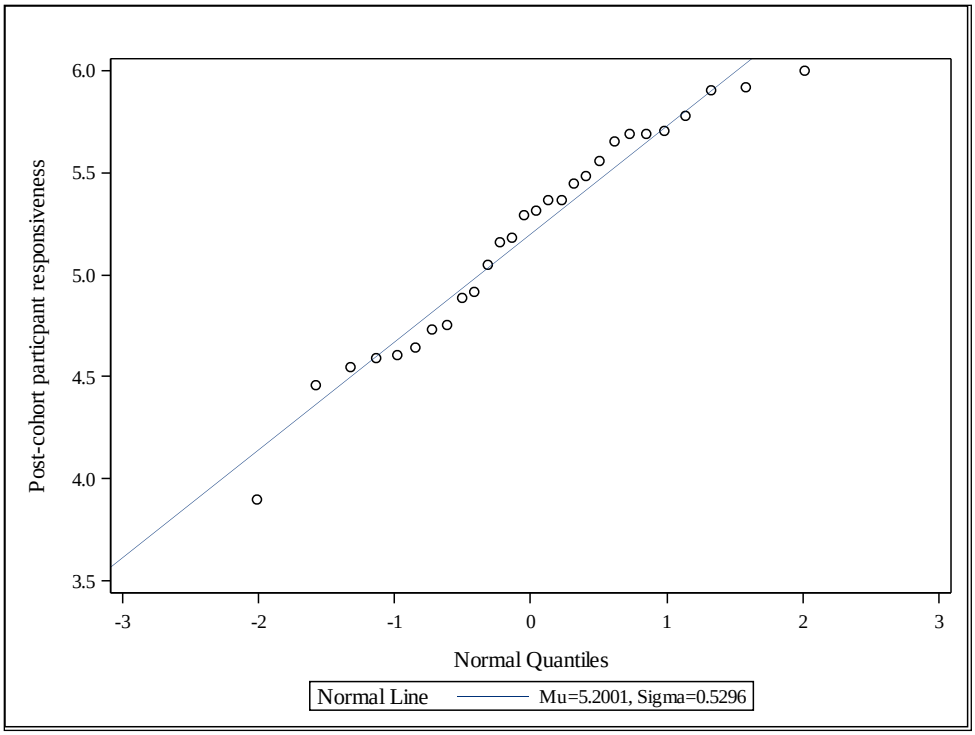

Figure 3.8. QQ plot for PCQ participant responsiveness scale.

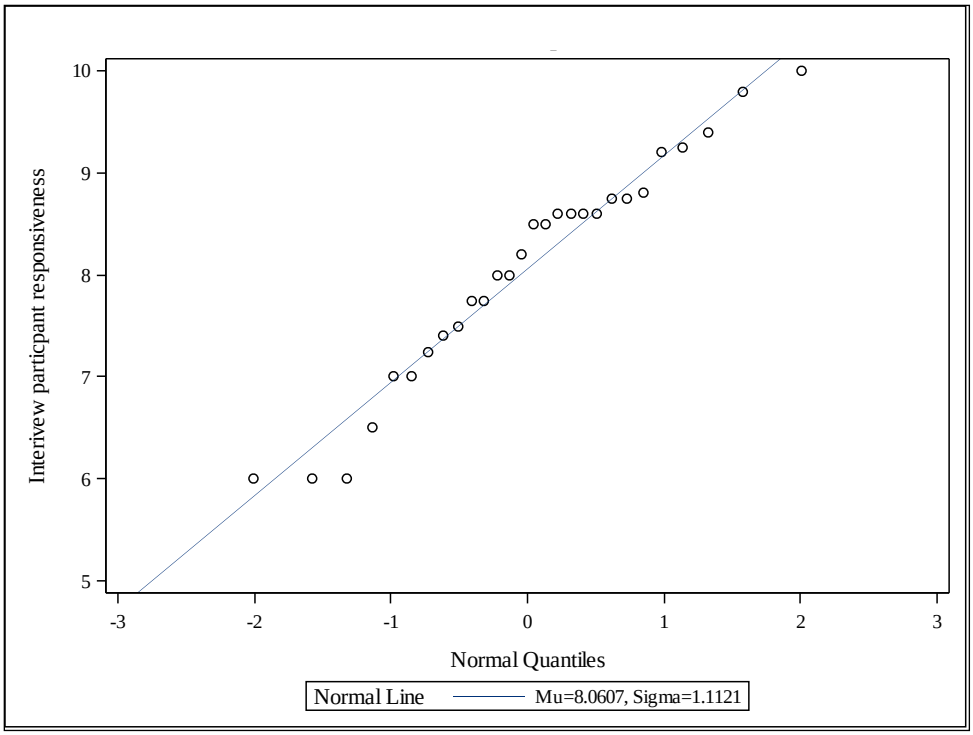

Figure 3.9. QQ plot for interview participant responsiveness scale.

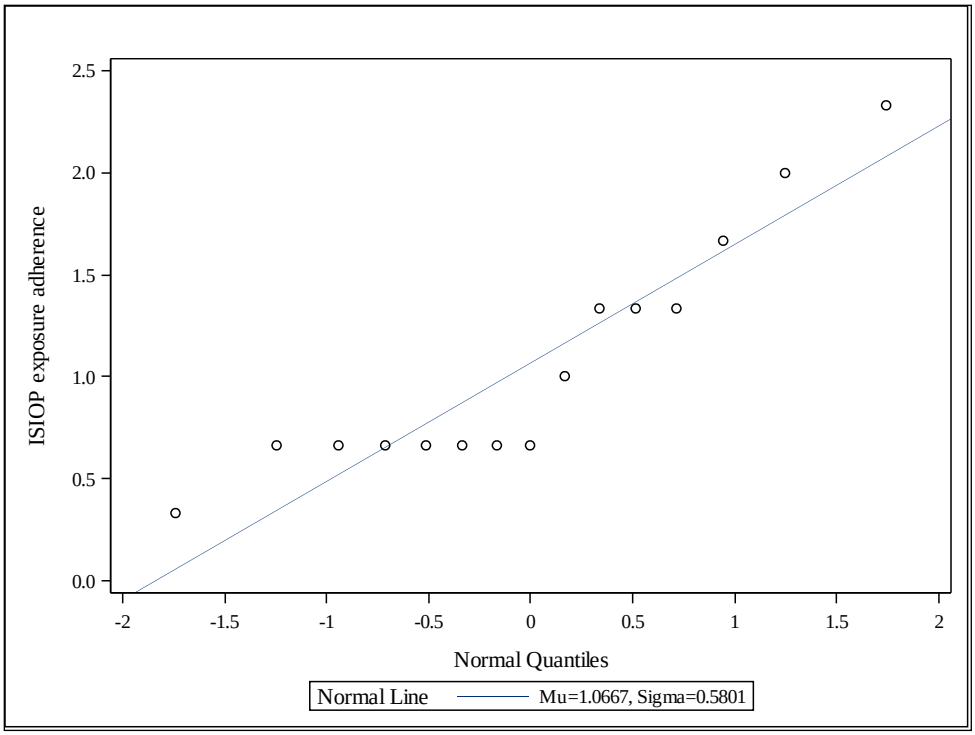

Figure 3.10. QQ plot for ISIOP exposure scale.

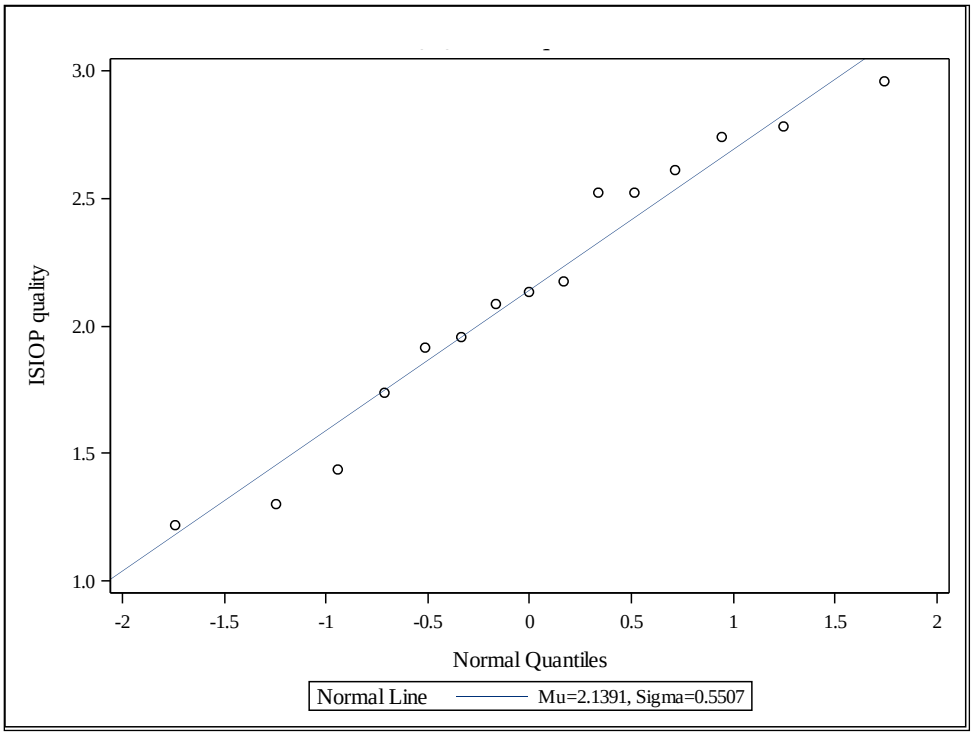

Figure 3.11. QQ plot for ISIOP quality scale.

CHAPTER 4

RESULTS

In this chapter I present the results of my study and answer my two research questions:

1. To what extent do the components of FOI as measured by instruments used in the TSI-A evaluation demonstrate convergent validity?
2. To what extent do the components of FOI as measured by instruments used in the TSI-A evaluation demonstrate discriminant validity?

To answer my research questions, I examined the correlations that I present in two MTMM matrices. The first MTMM matrix presents the correlations calculated from the results on all 28 teachers in a 7-by-7 matrix. These are the correlations among seven component scales represented on three instruments. The three instruments were used as part of the larger TSI-A evaluation and are the Target Activity Reflection (Reflection), the Post-Cohort Questionnaire (PCQ), and the Teacher Interview (Interview). The seven component scales are the Reflection adherence, Reflection quality, and Reflection participant responsiveness scales; the PCQ adherence, PCQ exposure, and PCQ participant responsiveness scales; and Interview participant responsiveness scale. The second MTMM matrix presents the results for the subset of 15 teachers who participated in the in-class observations in a 2-by-9 matrix. This matrix included the Inquiring into Science Instruction Observation Protocol (ISIOP), an instrument that I selected to supplement the FOI component data collected on the three TSI-A evaluation instruments. The matrix includes the correlations among the ISIOP exposure and the ISIOP quality scales and the other seven component scales from the first 7-by-7 matrix.

Ideally, an MTMM matrix presents the correlations among a number of components that are equally represented on two or more methods. This provides the ability to determine both convergent and discriminant validity among all the component scales. Given the focus of the

TSI-A evaluation, not all FOI components were equally represented across all instruments, resulting in some limitations to the conclusions I can make about the extent to which each FOI component is valid and distinct.

My analyses address the four criteria that Campbell and Fiske (1959) gave for establishing convergent and discriminant validity.

To provide evidence of convergent validity:

- The correlation among measurements of a single component on multiple methods (monocomponent-heteromethod comparisons) should be large and significantly different from zero.

To provide evidence for discriminant validity:

- The correlations among the multiple components measured on a single method (heterocomponent-monomethod comparisons) should be greater than the correlations among multiple components as measured with multiple methods (heterocomponent-heteromethod comparisons) but less than the correlations among the monocomponent-heteromethod comparisons (i.e., the comparisons used to examine convergent validity).
- The correlations among multiple components as measured using multiple methods (heterocomponent-heteromethod correlations) should be less than the correlations among the multiple components as measured using a single method (heterocomponent-monomethod comparisons)
- The pattern of component intercorrelations should be similar among the sets of heterocomponent-monomethod comparisons and the sets of the heterocomponent-heteromethod comparisons.

The third criterion addressing discriminant validity requires a matrix that includes a number of components equally represented across two or more methods. As a result of the unbalanced design of my matrices, I was unable to address the third criterion.

The monocomponent-monomethod comparisons provide evidence of the reliability of the FOI component scales. These are the correlations between a single component on a single method—that is, an estimate of the scales internal-consistency reliability. These values are expected to be the highest values in the MTMM matrix and provide the foundation for which to determine the convergent and discriminant validity.

In addition to the Campbell and Fiske (1959) criteria, I also use the expected correlation levels that were proposed by Salkind (2010). These correlation levels are used as a reference to determine the extent to which the FOI component scales produced correlation values that match the expected levels that support convergent and discriminant validity. I presented Salkind's theoretical figure in Chapter 3, Figure 3.1. The correlation levels that I defined to correspond to the actual correlation values are: low = $0 \leq r \geq .33$, moderate = $.34 \leq r \geq .66$, and high = $.67 \leq r \geq 1.0$.

For the remainder of this chapter I first present the results that show the FOI scales' reliability (monocomponent-monomethod comparisons). Second, I present the results for the monocomponent-heteromethod comparisons. These are the comparisons that help determine the FOI components' convergent validity and address my first research question. Third, I present the results of the heterocomponent-monomethod comparisons that are used to partially determine the FOI components' discriminant validity. Fourth, I present the results of the heterocomponent-heteromethod comparisons that also help determine the FOI components discriminant validity. The last two comparison results address my second research question. In each of the sections, I

present the results for the 7-by-7 matrix for all 28 teachers, followed by the results for the 2-by-9 matrix for the 15 teachers that were observed using the ISIOP. I conclude the chapter with a summary of the two matrices and indicate the extent to which each cell satisfied the expected correlation levels (Salkind, 2010). These final tables are used to help interpret my results that I discuss in Chapter 5.

FOI Scale Reliability (Monocomponent-Monomethod Comparisons)

To ensure that each of the FOI component scales has adequate internal constancy (i.e., $\alpha \geq .70$) (Nunnally & Bernstein, 1994), I calculated the Cronbach coefficient alpha for each. I first present the scales' internal consistency for the 7-by-7 matrix for all 28 teachers followed by the internal consistency results for only the 15 observed teachers.

All 28 Teachers

In Table 4.1, I present the 7-by-7 MTMM matrix for all 28 teachers. In Table 4.1 these internal consistency values are shown in parenthesis and are highlighted on the *reliability diagonal* (monocomponent-monomethod). These estimated values show good internal consistency among the different scales, which range from $\alpha = .76$ (Reflection adherence scale) to $\alpha = .93$ (Reflection participant responsiveness scale). The average internal consistency across all scales is $\alpha = .86$.

Table 4.1
7-by-7 MTMM Matrix for All 28 Teachers Showing Scale Reliability

		Reflection			Post Cohort			Interview
		A	Q	PR	A	E	PR	PR
Reflection	A	(.76)						
	Q	.50	(.83)					
	PR	.59	.74	(.93)				
PCQ	A	.16	.31	.41	(.91)			
	E	.22	.44	.24	.30	(.90)		
	PR	.51	.70	.76	.51	.52	(.86)	
Interview	PR	.28	.50	.54	.56	.31	.70	(.84)

A = adherence; E = exposure; Q = quality; PR = participant responsiveness

Only 15 Teachers

In Table 4.2, I present the 2-by-9 MTMM matrix for the 15 teachers who participated in the in-class observations. In Table 4.2, these internal consistency values are shown in parenthesis and highlighted. These two estimated values show that the ISIOP exposure scale had borderline good internal consistency with a value $\alpha = .69$. The ISIOP quality scale had high internal consistency shown by the value $\alpha = .93$. The average internal consistency across the two scales is $\alpha = .81$.

Results Addressing Research Question 1

To establish convergent validity of the FOI components—the degree to which components that should be related theoretically are related in reality—the correlation values for the monocomponent-heteromethod comparisons should be large and statistically significant from zero (Campbell & Fiske, 1959).

Monocomponent-Heteromethod Comparisons

All 28 teachers. In Table 4.3, I present the 7-by-7 MTMM matrix shown in Table 4.1 in which I have highlighted the cells containing the monocomponent-heteromethod comparisons. Only the FOI components of adherence and participant responsiveness were analyzed for

Table 4.2
2-by-9 MTMM Matrix for Only 15 Teachers Showing Scale Reliability

		ISIOP	
Instrument	FOI component	Exposure	Quality
ISIOP	Exposure	(.69)	
	Quality	- .45	(.93)
Reflection	Adherence	-.09	.10
	Quality	.18	-.07
	Participant Responsiveness	.22	.01
PCQ	Adherence	-.15	-.01
	Exposure	.04	-.35
	Participant Responsiveness	.18	-.19
Interview	Participant Responsiveness	.05	.07

Table 4.3

Monocomponent-Heteromethod Results for All 28 Teachers

		Reflection			Post Cohort			Interview
		A	Q	PR	A	Q	PR	A
Reflection	A	(.76)						
	Q	.50	(.83)					
	PR	.59	.74	(.93)				
PCQ	A	.16	.31	.41	(.91)			
	E	.22	.44	.24	.30	(.90)		
	PR	.51	.70	.76**	.51	.52	(.86)	
Interview	PR	.28	.50	.54**	.56	.31	.70**	(.84)

A = adherence; E = exposure; Q = quality; PR = participant responsiveness

** $p \leq .01$.

convergent validity in the 7-by-7 matrix. These were the only two FOI components for all 28 teachers that were measured using two or more methods.

There were two scales in my study that measured adherence: the Reflection adherence and the PCQ adherence scales. The correlation between these two scales is $r = .16$. This correlation value is neither large nor statistically significant from zero ($p = .43$) and thus does not meet the criterion for the convergent validity of adherence.

There were three scales in the 7-by-7 matrix that measured participant responsiveness: the Reflection participant responsiveness, PCQ participant responsiveness, and the Interview participant responsiveness scales. The correlation between the Reflection participant responsiveness and the PCQ participant responsiveness scales is $r = .76$, the correlation between the PCQ participant responsiveness and the Interview participant responsiveness scales is $r = .70$, and the correlation between the Reflection participant responsiveness and the Interview participant responsiveness scales is $r = .54$. The correlations between the Reflection participant responsiveness and the PCQ participant responsiveness scales and between the PCQ participant responsiveness and the Interview participant responsiveness scales were both large and statistically significant from zero ($p < .01$), thus meeting the criterion for the convergent validity

of participant responsiveness. The correlation between the Reflection participant responsiveness and the Interview participant responsiveness scales was statistically significant from zero ($p <$ .01) but produced a moderate correlation value—slightly lower than one that provides a high level of confidence in the convergent validity. Nevertheless, I interpret this value to be at a level that provides some evidence for convergent validity of participant responsiveness. The average correlation for all monocomponent-heteromethod comparisons involving participant responsiveness is $r = .67$—a borderline high correlation value.

Only 15 teachers. In Table 4.4, I present the 2-by-9 MTMM matrix shown in Table 4.2, with highlighted cells showing all the monocomponent-heteromethod comparisons. The FOI components of exposure and quality were analyzed for convergent validity in the 2-by-9 matrix. These were the two FOI components for the 15 observed teachers that were measured using two or more methods.

The two scales for measuring exposure were the PCQ exposure and the ISIOP exposure scales. The correlation between the PCQ exposure and the ISIOP exposure scales is $r = .04$. This correlation value clearly does not meet the criterion for the convergent validity of exposure.

Table 4.4
Monocomponent-Heteromethod Results for Only 15 Teachers

		ISIOP	
Instrument	FOI component	Exposure	Quality
ISIOP	Exposure	(.69)	
	Quality	- .45	(.93)
Reflection	Adherence	- .09	.10
	Quality	.18	- .07
	Participant Responsiveness	.22	.01
PCQ	Adherence	- .15	- .01
	Exposure	.04	- .35
	Participant Responsiveness	.18	- .19
Interview	Participant Responsiveness	.05	.07

The two scales in the 2-by-9 matrix that measured quality were the Reflection quality and the ISIOP quality scales. The correlation between the Reflection quality scale and the ISIOP quality scale is $r = -.07$, which obviously does not meet the criterion needed to provide evidence for the convergent validity of quality.

Results Addressing Research Question 2

To establish discriminant validity of the FOI components—the degree to which components that should *not* be related theoretically are *not* related in reality—the results for both the heterocomponent-monomethod and the heterocomponent-heteromethod comparisons were examined. To meet the first criterion for discriminant validity, the correlations between multiple components measured using a single method (i.e., heterocomponent-monomethod comparisons) should be smaller than the values needed to establish convergent validity (i.e., the monocomponent-heteromethod comparisons) (Campbell & Fiske, 1959). These values were expected to be at a moderate level (i.e., $.34 \leq r \geq .66$) (Salkind, 2010). To meet the second criterion for discriminant validity, the correlations between multiple components using multiple methods (i.e., heterocomponent-heteromethod comparisons) should be smaller than the correlations between multiple components using a single method (i.e., heterocomponent-monomethod comparisons) (Campbell & Fiske, 1959). These values were expected to be at a low level (i.e., $0 \leq r \geq .33$) (Salkind, 2010).

Heterocomponent-Monomethod Comparisons

All 28 teachers. In Table 4.5, I present the 7-by-7 MTMM matrix shown in Table 4.1, with highlighted cells showing all the heterocomponent-monomethod comparisons. In the 7-by-7 matrix these comparisons are for the Reflection and the PCQ—the two methods that addressed more than one component. The Interview only addressed participant responsiveness, so I could not conduct heterocomponent-monomethod analyses for this method.

Table 4.5

Heterocomponent-Monomethod Comparisons for All 28 Teachers

		Reflection			Post Cohort			Interview
		A	Q	PR	A	Q	PR	A
Reflection	A	(.76)						
	Q	.50	(.83)					
	PR	.59	.74	(.93)				
PCQ	A	.16	.31	.41	(.91)			
	E	.22	.44	.24	.30	(.90)		
	PR	.51	.70	.76	.51	.52	(.86)	
Interview	PR	.28	.50	.54	.56	.31	.70	(.84)

A = adherence; E = exposure; Q = quality; PR = participant responsiveness

Heterocomponent comparisons on the Reflection. The FOI components that were measured on the Reflection include adherence, quality, and participant responsiveness. As seen in Table 4.5, the correlation between adherence and quality is $r = .50$, the correlation between adherence and participant responsiveness is $r = .59$, and the correlation between quality and participant responsiveness is $r = .74$. The first two correlation values are at the expected moderate level (Salkind, 2010). The last correlation is larger than the moderate correlation level that is expected as evidence for discriminant validity. The average correlation among the FOI components on the Reflection is $r = .59$, which is smaller than the average correlations for the monocomponent-heteromethod comparisons and thus satisfies the second criterion needed to determine discriminant validity (Campbell & Fiske, 1959).

Heterocomponent comparisons on the PCQ. The FOI components that were measured on the PCQ include adherence, exposure, and participant responsiveness. The correlation between adherence and exposure is $r = .30$, between adherence and participant responsiveness is $r = .51$, and between exposure and participant responsiveness is $r = .52$. The first correlation value is lower than the expected moderate level (Salkind, 2010). The last two correlations are both at the expected moderate correlation level. The average correlation among the FOI components on the

PCQ is r = .44, which is smaller than the average correlation for the monocomponent-heteromethod comparisons and thus satisfies the second discriminant validity criterion (Campbell & Fiske, 1959).

The average correlation among all heterocomponent-monomethod (for both the Reflection and the PCQ) comparisons is r = .53—a moderate value that satisfies the expected moderate level (Salkind, 2010).

Only 15 teachers. In Table 4.6, I present the 2-by-9 MTMM matrix shown in Table 4.2, with highlighted cells showing the single heterocomponent-monomethod comparison. In the 2-by-9 matrix this is the comparison between exposure and quality on the ISIOP. The correlation between exposure and quality is r = - .45. Although this shows a negative relationship between the components, the magnitude of the correlation value is at the expected moderate level (i.e., .34 $\leq r \geq$ .66) (Salkind, 2010). However, this value is not lower than the monocomponent-heteromethod comparisons and does not satisfy the second criterion needed to support discriminant validity (Campbell & Fiske, 1959).

Table 4.6
Heterocomponent-Monomethod Comparisons for Only 15 Teachers

		ISIOP	
Instrument	FOI component	Exposure	Quality
ISIOP	Exposure	(.69)	
	Quality	- .45	(.93)
Reflection	Adherence	- .09	.10
	Quality	.18	- .07
	Participant Responsiveness	.22	.01
PCQ	Adherence	- .15	- .01
	Exposure	.04	- .35
	Participant Responsiveness	.18	- .19
Interview	Participant Responsiveness	.05	.07

Heterocomponent-Heteromethod Comparisons

All 28 teachers. In Table 4.7, I present the 7-by-7 MTMM matrix shown in Table 4, with highlighted cells showing the heterocomponent-heteromethod comparisons. To satisfy the second criterion for discriminant validity, these values should be lower than the heterocomponent-monomethod comparisons. These comparisons are expected to be at a low correlation level (i.e., $0 \leq r \geq .33$). For clarity of presentation, I have categorized the *r* values into the three correlation level groups (i.e., 0 to .33, .34 to .66, and .67 to 1.0). I present these comparisons in ascending *r* value order in Table 4.8. There were 11 heterocomponent-heteromethod comparisons in the 7-by-7 matrix. The results that I present in Table 4.8 show that of the 11 comparisons, only 5 are at the expected low correlation level. The results show that of the remaining six comparisons, five are at the moderate level and one is at the high level. The average correlation among all the heterocomponent-heteromethod comparisons is $r = .41$. Overall, the average correlation is smaller than the average correlation among the heterocomponent-monomethod comparisons, which satisfies the second criterion needed to determine discriminant validity but was higher than the ideal low correlation value (i.e., ≤ .33).

Table 4.7
Heterocomponent-Heteromethod Comparisons for All 28 Teachers

		Reflection			Post Cohort			Interview
		A	Q	PR	A	Q	PR	A
Reflection	A	(.76)						
	Q	.50	(.83)					
	PR	.59	.74	(.93)				
PCQ	A	.16	.31	.41	(.91)			
	E	.22	.44	.24	.30	(.90)		
	PR	.51	.70	.76	.51	.52	(.86)	
Interview	PR	.28	.50	.54	.56	.31	.70	(.84)

A = adherence; E = exposure; Q = quality; PR = participant responsiveness

Table 4.8
Results from the Heterocomponent-Heteromethod Comparisons in Ascending (r) Value Order for all 28 Teachers

Comparisons	Category (Pearson *r)*		
	0 to .33	.34 to .66	.67 to 1.0
R adherence vs. PCQ exposure	.22		
R participant responsiveness vs. PCQ exposure	.24		
R adherence vs. I participant responsiveness	.28		
PCQ exposure vs. I participant responsiveness	.31		
R quality vs. PCQ adherence	.31		
R participant responsiveness vs. PCQ adherence		.41	
R quality vs. PCQ exposure		.44	
R adherence vs. PCQ participant responsiveness		.51	
R quality vs. I participant responsiveness		.50	
PCQ adherence vs. I participant responsiveness		.56	
R quality vs. PCQ participant responsiveness			.70

Only 15 teachers. In Table 4.9, I present the 2-by-9 MTMM matrix shown in Table 4.2, with highlighted cells showing all the heterocomponent-heteromethod comparisons. For clarity of presentation, I have categorized the *r* values into the three correlation level groups (i.e., 0 to .33, .34 to .66, and .67 to 1.0). I present these comparisons in ascending *r* value order in Table 4.10. (I used the absolute value of the correlations for sorting purposes.) There were 12 heterocomponent-heteromethod comparisons in the 7-by-7 matrix. The results that I present in Table 4.10 show that of the 12 comparisons, all but 1 are at the expected low correlation level.

Table 4.9
Heterocomponent-Heteromethod Comparisons for Only 15 Teachers

		ISIOP	
Instrument	FOI component	Exposure	Quality
ISIOP	Exposure	(.69)	
	Quality	- .45	(.93)
Reflection	Adherence	-.09	.10
	Quality	.18	-.07
	Participant Responsiveness	.22	.01
PCQ	Adherence	-.15	-.01
	Exposure	.04	-.35
	Participant Responsiveness	.18	-.19
Interview	Participant Responsiveness	.05	.07

Table 4.10
Results from the Heterocomponent-Heteromethod Comparisons in Ascending (r) Value Order for Only the 15 Teachers

Comparisons	Category (Pearson *r)*		
	0 to .33	.34 to .66	.67 to 1.0
ISIOP quality vs. R participant responsiveness	.01		
ISIOP quality vs. PCQ adherence	- .01		
ISIOP exposure vs. I participant responsiveness	.05		
ISIOP quality vs. I participant responsiveness	.07		
ISIOP exposure vs. R adherence	-.09		
ISIOP quality vs. R adherence	.10		
ISIOP exposure vs. PCQ adherence	-.15		
ISIOP exposure vs. R quality	.18		
ISIOP exposure vs. PCQ participant responsiveness	.18		
ISIOP quality vs. PCQ participant responsiveness	-.19		
ISIOP exposure vs. R participant responsiveness	.22		
ISIOP quality vs. PCQ exposure		-.35	

The overall correlation (absolute) value among all heterocomponent-heteromethod comparisons is $r = .13$. Overall, the average magnitude of the correlation is smaller than the heterocomponent-heteromethod comparison, which satisfies the second criterion needed to determine discriminant validity.

Summary of Findings About Convergent and Discriminant Validity

In Table 4.11, I present a repeat of the 7-by-7 matrix. In this table I provide the actual correlations values followed by the expected correlation levels (indicated by L (low), M (moderate), and H (high)). I have highlighted the cells that did *not* match the expected correlation levels among the comparisons. In Table 4.12, I present a repeat of the 2-by-9 matrix. In this table, I also highlighted the cells where the correlation values did not match the expected correlation levels. I will discuss the extent to which the different comparisons support convergent and discriminant validly for the FOI components in Chapter 5.

Table 4.11

7-by-7 MTMM Matrix Results Showing the Extent to Which the Correlation Values Matched the Expected Levels for all 28 Teachers

		Reflection			Post-Cohort Questionnaire			Interview
		A	Q	PR	A	Q	PR	A
Reflection	A	(.76)/H						
	Q	.50/M	(.83)/H					
	PR	.59/M	.74/M	(.93)/H				
PCQ	A	.16/H	.31/L	.41/L	(.91)/H			
	E	.22/L	.44/L	.24/L	.30/M	(.90)/H		
	PR	.51/L	.70/L	.76/H	.51/M	.52/M	(.86)/H	
Interview	PR	.28/L	.50/L	.54/H	.56/L	.31/L	.70/H	(.84)/H

A = adherence; E = exposure; Q = quality; PR = participant responsiveness

Expected correlation level: L = low (0 - .33); M = moderate (.34 - .66); H = high (.67 – 1.0)

Table 4.12

2-by-9 MTMM Matrix Results Showing the Extent to Which the Correlation Values Matched the Expected Levels for Only 15 Teachers

		ISIOP	
Instrument	FOI component	Exposure	Quality
ISIOP	Exposure	(.69)/H	
	Quality	- .45/M	(.93)/H
Reflection	Adherence	-.09/L	.10/L
	Quality	.18/L	-.07/M
	Participant Responsiveness	.22/L	.01/L
PCQ	Adherence	-.15/L	-.01/L
	Exposure	.04/M	-.35/L
	Participant Responsiveness	.18/L	-.19/L
Interview	Participant Responsiveness	.05/L	.07/L

Expected correlation level: L = low (0 - .33); M = moderate (.34 - .66); H = high (.67 – 1.0)

CHAPTER 5

DISCUSSION

Program evaluators and researchers have emphasized the need to better understand the relationship among the FOI components (e.g., Carroll et al., 2007). However, a lack of consensus about how best to define FOI and its components, as well as how best to collect FOI data, continue to be a challenge. This limits the extent to which researchers and evaluators have a comprehensive understanding of the relationships among the FOI components and of how FOI affects program effectiveness. Improving our understanding about the relationships among the FOI components, as well as of the validity of the inferences from data collected about the components, will allow researchers and evaluators to better determine how to focus their efforts when collecting FOI data. Understanding the relationship among the FOI components also has implications for how we conceptualize the FOI components. Accordingly, the purpose of my study was to use an MTMM matrix (Campbell & Fiske, 1959) for systematically examining the convergent and discriminant validity of four components of FOI (adherence, exposure, quality, and participant responsiveness) as they were operationalized in a single evaluation study.

Using an MTMM matrix, I examined the convergent and discriminant validity of the FOI components to answer my two research questions.

1. To what extent do the components of FOI as measured by instruments used in the TSI-A evaluation demonstrate convergent validity?
2. To what extent do the components of FOI as measured by instruments used in the TSI-A evaluation demonstrate discriminant validity?

The data that I examined were collected with four instruments—three self-report instruments (the Reflection, PCQ, and Interview) collected from 28 O‘ahu and Kaua‘i teachers, and an observation instrument (the ISIOP) for collecting data from only the 15 O‘ahu teachers.

I begin this chapter with a discussion about the ISIOP data and their usefulness for examining validity. For the remainder of the chapter, I discuss the results as they address each of the two research questions.

Discussion of the ISIOP Results

Despite my best intentions, the ISIOP data contributed inconclusively to my study. In this section, I explain why and provide some conclusions about conducting observations to examine FOI.

I chose to use observations as an alternative method for collecting FOI data to combat a potential self-reporting bias (Hansen and McNeal, 1999) of the three other instruments in my study. I selected the ISIOP due to its extensive development and attention to aspects that are associated with quality inquiry-based instruction (Minner & DeLisi, 2012). The instrument was designed to comprehensively capture the nature of the teachers' support for scientific practices such as questioning and assessment strategies—aspects of the TSI-A model. Given the high level of interrater reliability achieved between me and the second rater, both during the training to use the instrument and during its use in the TSI-A classrooms, we had evidence that we were accurately measuring teachers. However, the MTMM matrices show minimal evidence supporting convergent and discriminant validity—findings for which I suggest five interpretations.

Inadequate Alignment of the ISIOP with TSI-A

The ISIOP might not have been sufficiently aligned with the unique properties of the TSI-A project. Teacher behaviors were coded for general inquiry-based science instruction practices, some of which were not directly comparable to the behaviors specific to the TSI-A project (i.e., the extent to which teachers address the phases and modes of inquiry), a manifestation of content underrepresentation or construct-irrelevant variance (Cook & Campbell, 1979; Messick, 1989).

Conducting Observations by Program Experts

Perhaps the ISIOP should only have been used by TSI-A program experts. Such experts could translate the behaviors that were shown on the ISIOP to what the behaviors might mean in the context of the TSI-A language. For example, program-expert observers could address how the ISIOP-coded behaviors best represented the extent to which the teachers were implementing the TSI-A phases and modes of inquiry.

Problems in Measuring Quality

Measuring program quality in observations might be too difficult to assess by anyone other than a subject matter expert. Observing and coding quality is a complex judgment task that is often more difficult than measuring other aspects of program fidelity. Sanchez et al. (2007) concluded that to produce a good measure of quality, observers should be skilled in the intervention or extensively trained in the observation protocol. I go beyond that, however, and conclude that observers should be skilled in the intervention *and* extensively trained in the observation protocol to produce a good measure of quality. This is supported by research that found a good relationship among expert ratings on instruments that were designed specifically for the study (Brandon et al., 2007).

Ruiz-Primo (2006) notes that some dimensions of program components such as a participant's understanding of content are difficult to observe; quality may also be one such component. More research is needed to better understand how, and if, quality can be accurately measured.

Measuring Exposure With the ISIOP

TSI-A exposure might not have been appropriately measured on the ISIOP. To measure exposure, the ISIOP used Likert scales to summarize the extent to which the teachers addressed the different aspects of inquiry-based science teaching. Exposure refers to the *amount* of the

program content that is delivered to students, as well as the *frequency* in which specific program techniques are implemented (Dane & Schneider, 1998). These features cannot be accurately assessed as an ordinal variable. One may be able to say that there was a low or high level of exposure, but conclusions about levels between these extremes might be unduly arbitrary (Durlak & Dupre, 2008). Moreover, the scale points used to summarize exposure with the ISIOP had highly unequal distances between the points (i.e., 0 (0%), 1 (1% to 10%), 2 (11% to 50%), and 3 (51% to 100%)). This is reflected by the non-normal distribution of the ISIOP exposure scale and provides additional support that exposure might have been inappropriately measured on the ISIOP. The limitation of only being able to sample a single lesson may also have been a source of instability in measuring exposure.

A Potential for Method Effects

It is beyond the scope of this study to go into depth about the benefits of observations over self-reports, but clearly the potential exists for a bias in self-reports that is missing in observations. Thus, the lack of correlation between the ISIOP and the Reflection and the PCQ might show strong method effects that reflect a self-report bias in the latter two instruments. Some evidence for this is found in the results of the heterocomponent-monomethod and heterocomponent-heteromethod analyses used to show evidence of discriminant validity. The average of the 12 correlations calculated of these analyses was $r = .13$—a level that failed to establish any relationship between the scales or the components they were intended to measure.

Besides acknowledging this possibility, it is impossible to arrive at a strong conclusion about the degree to which the ISIOP taps aspects of implementation that the other instruments do not. My best speculation is that, more than reflecting a method effect, the virtually null correlations between the ISIOP and the self-report instruments reflect potential problems in ISIOP measurements within the context of TSI-A.

In conclusion, the potential mismatch of the ISIOP and TSI-A points out the need for careful delineation of what comprises an appropriate FOI instrument in a particular setting for a particular purpose. It also highlights the importance of providing an explicit definition of quality within the study's context, as well as ensuring that each FOI component is measured in accordance to its definition (e.g., measuring exposure in units of frequency or amount).

Discussion of the Results for the PCQ, Reflection, and Interview

For the remainder of this chapter I present my conclusions and interpretations in the same order in which I presented my results in Chapter 4. I begin with the conclusions about the FOI component scales' reliability (monocomponent-monomethod comparisons). Next, I discuss the convergent validity findings shown in the results of the monocomponent-heteromethod comparisons and address my first research question. Finally, the last two sections address the discriminant validity findings and address my second research question. I conclude this chapter with a summary of the findings, suggestions for future research, and comments about my study's limitations.

FOI Scale Reliability (Monocomponent-Monomethod Comparisons)

The results for all the monocomponent-monomethod correlations show that there was a high level of internal-consistency reliability across all FOI component scales. The scales are the highest values in the matrix and satisfy the expected correlation levels, thus establishing a solid foundation for examining the convergent and discriminant validity of the operationalized FOI components using the MTMM approach.

The technical quality of an instrument, including its reliability, is an aspect of content validity (Messick, 1989). In addition to the high reliability of the FOI scales, the concerted effort of the evaluation team to consider and include aspects of FOI during the instruments' development provides evidence of content validity. The evaluation team's systematic review of

all the instrument items and careful assignment of the items to the FOI components provides additional support for content validity. This systematic review also limits the potential systematic measurement error that might be an issue if such careful construction of the instruments and careful item assignment had not been done. A caveat is that the elimination of items with low item-total correlation might have resulted in construct underrepresentation to an unknown degree.

Discussion of the Results Addressing Research Question 1

Monocomponent-heteromethod comparisons. The values in an MTMM matrix that address convergent validity are the correlations for a component as it is measured on two or more instruments. These values are expected to be lower than the scales' internal consistency reliability coefficients and have higher *r* values than the remaining MTMM matrix correlations. To establish evidence of convergent validity, the FOI components should be more strongly related across multiple instruments than they should be with other components measured on a single instrument or with other components measured on multiple instruments.

The MTMM matrix for the 28 teachers shows the degree of convergent validity for two components—participant responsiveness and adherence. Because exposure and quality were only measured on a single instrument (PCQ and Reflection, respectively), I was unable to conduct convergent validity analyses for these two components.

Overall, there is a high level of confidence for the convergent validity of participant responsiveness as measured across the three instruments. The convergent validity of participant responsiveness as measured on the Interview and the Reflection showed a somewhat lower-than-desirable correlation (r = .54—a moderate value when it should be high), but I interpret this to be at a high enough level to provide some evidence of convergent validity. These results show that teacher accounts of their students' receptivity and learning of the TSI-A content and scientific

procedures, as well as their perceived value of, and success in using, the TSI-A pedagogy, are consistent across methods. The results suggest that participant responsiveness can be validly measured in multiple ways.

In contrast, the results show a lack of convergent validity for adherence. I speculate three possible general reasons for these findings. First, they might be due in part to the teachers' dichotomous conceptualization of adherence. I define adherence as the extent to which the teachers implemented the steps in the target activities. In its simplest form, it can be assessed by asking questions such as, "In implementing the program, did you or did you not follow all the procedures you were taught?" or "From the following list of steps, which ones did you follow?" However, on both the Reflection and the PCQ, adherence was treated as an ordinal variable (i.e., answered with a Likert scale). That is, on both instruments the response options ranged from *not at all* (1) to *very much* (5) on the Reflection, and *not at all* (1) to *completely* (6) on the PCQ. Except from memory, it was not specified how the teachers could know how many TSI-A steps needed to be implemented for them to provide moderate ratings. That is, the percentage of TSI-A steps that represent ratings of 2, 3, or 4 most likely was unclear to the teachers. Consequently, I speculate that adherence should be treated as a dichotomous variable. Preferably, respondents should be given checklist asking whether they implemented each step (i.e., "I only covered steps a, b, and c but not x, y, or z"). Indeed, research shows that adherence data are often gathered with the use of checklists, logs, and staff observations and reported by researchers and evaluators as the proportion of the components addressed (e.g., Frank, Bose, & Schrobenhauser-Clonan, 2014; Metz et al., 2013; Noell, Witt, Gilbertson, Ranier, & Freeland, 1997; Resnicow et al., 1998).

Second, in the Interview conducted at the end of the program, several teachers expressed that the program was too complex to determine the extent to which the TSI-A pedagogy (i.e., phases,

modes, etc.) were explicitly addressed with their students (Seraphin, 2014). This suggests that teachers' perceptions about the extent to which they explicitly addressed the necessary steps were not uniformly clear. The lack of clarity about the extent to which the TSI-A components were implemented may have led to inaccuracies in their self-reporting of adherence. This highlights the importance of ensuring that teachers have a clear understanding of the components they are implementing and what differential adherence looks like in a classroom setting.

Third, the Reflection was administered repeatedly over the final year of the project, with each teacher completing at least 12 of the 15 Reflections immediately following their implementation (in contrast to the single administration of the PCQ after project activities concluded). Also, each of the central TSI-A components were divided into separate items on the Reflection, which also added to the overall item pool. This approach reflects my speculation about the specificity to all the steps of the program that must be provided in an instrument. Measurement error may also have been minimized simply by having more responses on the Reflection adherence scale (Dillman, 2007). In contrast, there is only one general item on the PCQ that addressed adherence (i.e., to what extent did you implement all the steps of the target activities in each of the following modules?), which was separated into four sub-items (i.e., for Module 1, Module 2, etc.). Taking into account all the steps and aspects of the TSI-A model (e.g., five phases, ten modes, and so on), I suggest that the PCQ adherence item missed many of the nuances of the project components. Also, because of only having one global item that was intended to cover multiple aspects of the TSI-A project, teachers may have responded to the items in terms of how much they generally favored the entire project (i.e., halo effect).

Summarizing my conclusions about convergent validity, the results show:

- Participant responsiveness has a high level of convergent validity as measured across three different instruments.
- The correlation between the Interview and Reflection participant responsiveness scales produced a lower-than expected moderate correlation, but it is at an acceptable level to support convergent validity of participant responsiveness between these scales.
- Adherence has a low level of convergent validity as measured across two different instruments. The high degree of discrimination between these two scales, indicated by the low correlation between the scales (i.e., $r = .16$), suggests that they are measuring different components.
- It may be inappropriate to treat the FOI component adherence as an ordinal variable (i.e., using a Likert scale for a component that is dichotomous in nature).
- Due to the complexity of the TSI-A components, teachers may not have had the level of understanding to accurately self-report the extent to which they followed all the steps or procedures in an activity. This may be particularly true on the PCQ, which was administered at the end of the school year, and not only required understanding but also a clear recollection about the degree to which they implemented all the steps. The PCQ may have provided results that are not a true measure of adherence to project components.
- The more consistent results among the Reflection adherence comparisons may be the result of a larger number of items that minimized measurement error.
- Overall, the results suggest that the instruments used to evaluate the TSI-A provide a valid measure of the teachers' participant responsiveness to project activities. Whereas,

the results suggest that adherence was not accurately measured using the TSI-A instruments.

Discussion of the Results Addressing Research Question 2

Campbell & Fiske (1959) listed three criteria that provide evidence of discriminant validity. Due to the unbalanced design of my matrix, I only address the first and the second criterion.

Heterocomponent-monomethod comparisons. To fit the expected correlation levels, correlations among multiple components measured with a single method are expected to be moderate values and to be lower than the correlation values produced by a single component measured with multiple methods (Campbell & Fiske, 1959; Salkind, 2010). The data collected in the TSI-A study only allowed for heterocomponent-monomethod comparisons on the Reflection and the PCQ—the two instruments that collected data on more than one FOI component.

Heterocomponent comparisons on the Reflection. The results of the comparisons among the FOI components measured on the Reflection show acceptable levels of discrimination between the FOI components of quality and adherence and between participant responsiveness and adherence. These results do not provide unambiguous evidence that the scales are providing valid measures of the FOI components but lend support to the notion that they are measuring different components.

The higher-than-expected correlation between quality and participant responsiveness on the Reflection (.74, high when it should be moderate) fails to show discrimination between these two FOI components. A closer examination of the comparisons between the Reflection quality scale and the two other participant responsiveness scales shows a similarly high correlation between the Reflection quality scale and the PCQ participant responsiveness scale (.70, high when it should be low), as well as a higher-than-expected moderate correlation between the Reflection quality scale and the Interview participant responsiveness scale (.50, moderate when it should be

low). Altogether, this suggests that quality shares a closer relationship to participant responsiveness, at least within the context of TSI-A. That is, self-ratings of quality (i.e., *how well* it was implemented) might reflect teachers' perceptions about the extent to which the activities positively affected student learning.

Heterocomponent comparisons on the PCQ. The results of the comparisons among the FOI components measured on the PCQ shows evidence for discrimination between the FOI components of participant responsiveness and adherence and between participant responsiveness and exposure. However, the comparison between adherence and exposure was lower than the expected moderate correlation needed to show discrimination between components using the same method. This correlation level suggests that there is discrimination between exposure and adherence due to both component and method differences, which mostly likely can be explained by the systematic measurement error on the PCQ adherence scale. That is, because the internal-consistency reliability was high for this global scale of adherence ($\alpha = .91$), it may suggest that teachers were probably responding with a general impression of how much they subscribed to the entire program (i.e., a halo effect). It may also be the result in the teachers' belief that because they were *expected* to follow the prescribed steps in their implementation of the target activities, they *must have* followed the prescribed steps as intended (i.e., subject-expectancy effect).

The moderate correlation between adherence and participant responsiveness on the PCQ reflects a possible similarity in teachers' perceptions between adherence and participant responsiveness. To better examine this proposed relationship between PCQ adherence and participant responsiveness, I examined the comparisons between the PCQ adherence scale and the Reflection participant responsiveness scale and between the PCQ adherence scale and the

Interview Participant responsiveness scale. There was a higher-than-expected correlation between both the PCQ adherence scale and the Reflection participant responsiveness scale (.41, moderate when it should be low), and the PCQ adherence scale and the Interview Participant Responsiveness scale (.56, moderate when it should be low). The results from these comparisons suggest that there was a higher-than-expected relationship between PCQ adherence scale and the participant responsiveness scales. Perhaps to the teachers, they had adhered to a program when they saw positive effects on their students' learning. By virtue of their training and experience, teachers may have been more focused on the extent to which their students were learning the content rather than the extent to which they followed the steps and procedures of a program or curriculum that they were implementing. Also, research on professional development has shown that teacher' attitudes and beliefs about the value of a project are related to the extent to which they observe improvements in student learning (Guskey, 2002). Therefore, the teachers' ratings of adherence on the PCQ might reflect somewhat their recollection about the extent to which they perceived the TSI-A components to be of value in improving student learning. Alternatively, however, this might also be explained as a result of systematic measurement error on the PCQ adherence scale, which is overinflating the strength of the relationship.

Heterocomponent-heteromethod comparisons. To satisfy Campbell and Fiske's second criterion for discriminant validity, the heterocomponent-heteromethod correlations are expected to be lowest values in the matrix. These are the correlations between different components as measured by different instruments. Low correlations in an MTMM matrix provide evidence that there is both good component and method discrimination among the scales.

In my matrix, there are a total of 11 heterocomponent-heteromethod comparisons. The results show that five of these satisfied the expected low correlation levels needed to support

discriminant validity. The remaining six high and moderate correlations require further discussion. These unexpected moderate and high correlations among the heterocomponent-heteromethod correlations support some of my earlier arguments about the relationships among the FOI component scales. I discuss the unexpected high and moderate correlations as follows:

- The unexpected high correlation between the Reflection quality scale and the PCQ participant responsiveness scale (.70, high when it should be low), as well as the higher-than-expected moderate correlation between the Reflection quality scale and Interview participant responsiveness scale (.50, moderate when it should be low) supports a relationship between quality and participant responsiveness within the context of TSI-A.
- The correlation between the PCQ adherence scale and the Interview participant responsiveness scale (r = .56—a moderate value when it should be low), as well as the correlation between the PCQ adherence scale and the Reflection participant responsiveness scale (r = .41—also a moderate value when it should be low), provides evidence that there is a closer-than-expected relationship between the PCQ adherence scale and participant responsiveness. I interpret this relationship with caution due to the potentially high degree of potential biases on the PCQ adherence scale (i.e., halo effect, subject-expectancy effect, or social desirability bias), which may have overinflated the correlations that show a relationship between these components
- The higher-than-expected moderate correlations between the Reflection adherence scale and the PCQ participant responsiveness scale and between the PCQ exposure scale and the Reflection quality scale do not follow any expected pattern that is needed to provide evidence of discriminant validity between the scales. Potentially, these higher-than-expected correlations may be the results of systematic measurement error on these scales,

which resulted in inflated correlation values. That is, given the length of the PCQ (i.e., 100 items), and given that it was administered at the conclusion of the project, teachers were probably *satisficing* on their responses (Krosnick, 1991). That is, the teachers might have responded to the items by providing ratings to meet the need of simply completing the long questionnaire after a long PD, rather than providing ratings that optimally represented their exposure, adherence, or participant responsiveness.

- The heterocomponent-heteromethod comparison results provide some evidence of discrimination for the FOI component exposure. It consistently produced the expected low correlation levels when compared to the two participant responsiveness scales (Interview and Reflection), and also satisfied the expected low correlation level when compared with Refection adherence. It failed to satisfy the expected levels of correlation with PCQ adherence and Reflection quality, however. I interpret this to reflect method effects on the Reflection quality scale, as well due to the potential systematic measurement error on the PCQ adherence scale. Apart from these two comparisons, the results, overall suggest that exposure is a distinct FOI component with the context of TSI-A.

Summarizing my conclusions about discriminant validity based on the heterocomponent-monomethod and the heterocomponent-heteromethod comparisons, the result show:

- Exposure as measured on the PCQ showed overall good discrimination when compared to the other FOI scales. Evidence for this is provided in both the heterocomponent-monomethod and heterocomponent-heteromethod results.
- Adherence as measured on the Reflection provides evidence that it is a distinct component. (Because convergent validity was not established for adherence, there is not

enough evidence to either confirm or reject that the Reflection provided a valid measure of adherence).

- Adherence as measured on the PCQ has a closer relationship to all three participant responsiveness scales and therefore is not a distinct component. The potentially high level of systematic measurement error on the PCQ adherence scale may have overinflated or underinflated any potential relationships or non-relationships with the other FOI components.
- Quality as measured on the Reflection has a closer relationship to participant responsiveness and is therefore not a distinct component within the context of TSI-A.
- Participant responsiveness had mixed evidence of discriminant validity seen in both the heterocomponent-monomethod and the heterocomponent-heteromethod comparisons. It showed a closer than expected relationship with PCQ adherence and the Reflection quality scales but did discriminate well with Reflection adherence and PCQ exposure.
- Thee mixed evidence of discriminant validity for participant responsiveness may have been due to the multidimensionality of the component within my study's context. That is, participant responsiveness was addressed on items that asked about perceived value, perceived success, and perceived effects on student learning and engagement, which may have resulted in shared variance with the other components.

Summary, Lessons Learned, and Future Research

Overall, the pattern of correlations in the 7-by-7 matrix follows the hierarchy of criteria specified by Campbell and Fiske (1959) that provides evidence of both convergent and discriminant validity. However, due to the unbalanced design of my MTMM matrix, the strength of my conclusions about the extent to which the FOI components are valid and distinct is

somewhat incomplete. That is, evidence for convergent validity was limited to only adherence and participant responsiveness, and there were unexpected relationships that could only be explained by systematic measurement error or potential method effects. Regardless, I have several conclusions that can inform future FOI research.

The evidence showed that participant responsiveness had high convergence across three instruments, which suggests that evaluators can accurately capture participant responsiveness (a) using a self-report measure immediately following teachers' implementation of an activity (e.g., Reflection); (b) on an extensive self-report end-of-project questionnaire covering multiple aspects of a project (e.g., PCQ); and (c) during computer-assisted face-to-face interviews in which respondents are expected to explain their self-reported ratings (e.g., Interview). This has particular implications for the efficiency of collecting data on participant responsiveness as it was operationalized in this study.

Meta-analytic reviews of the FOI literature have shown that participant responsiveness is vastly underreported in FOI-related research (e.g., Dane and Schneider, 1998; Durlak and Dupre, 2008; Gould et al., 2016; Mihalic, 2004. The mixed discriminant-validity evidence for participant responsiveness found in this study suggests that future research might examine the extent to which the multidimensionality of definitions of participant responsiveness affect convergence and discrimination. Is participant responsiveness a measure of participants' enthusiasm and level of participation (Berkel, 2010), perceived student learning (Lynch, & O'Donnell, 2005), interest in a program (Durlak & Dupre, 2008), engagement in project activities (Mihalic, 2004), or acceptance of an intervention's content (Ibrahim & Sadini, 2016)?

The evidence that adherence had low convergence across two instruments suggests that either one or both of the instruments provided a poor measure of adherence. I reiterate that to get an

accurate measure of adherence, data should be collected in a manner that can provide a proportion of steps or procedures addressed—assuming that is an important indicator of project success—rather than using a Likert scale.

The closer-than-expected relationship between the PCQ adherence scale and the participant responsiveness scales provides some evidence that teachers may have provided a global value rating of the project when rating their adherence. Teachers might simply be unable to accurately rate the extent that they covered all the steps of all project activities. Furthermore, having them retrospectively rate the extent that they covered all steps of activities that happened months earlier might be unrealistic. I speculate that teachers might provide ratings closely reflecting the perceived overall value of projects. The notion of approaching a developer's theoretical ideal of how teachers should implement what they learned in PD is elusive and might not be routinely measurable.

The closer-than-expected relationship between the Reflection quality scale and the participant responsiveness scales supports my speculation and suggest that quality may be conceptualized in terms of teachers' perceptions about the value they placed on the project or the extent that it engaged their students and improved student learning. This conclusion is consistent with PD research that suggests teachers will use what they learned in PD to the extent that they receive regular positive feedback about student performance (Guskey, 2002). Future research might examine the extent to which a measure of quality is feasible in the evaluation of an innovative PD project.

Much is yet to be learned about FOI components within the context of education programs. Therefore, to enhance our understanding about the nature of FOI, studies like mine should be repeated across many types of projects. Future FOI studies must also ensure that each component

is equally and comprehensively addressed. Not having each component equally and comprehensively addressed severely limits the extent to which convergent and discriminant validity can be examined. Future studies must also ensure that that each component of FOI is well defined and appropriately measured based on the nature of the component.

Limitations

There are several limitations to this study. First, there is a small sample size of teachers in my study (*N*=28). The small sample size has the potential to decrease the power of the study. However, tests of normality suggest that the results were not extremely affected by the small sample sizes but it did limit the types of analyses I was able to conduct. Second, eliminating items with low item-total correlations may have resulted in a degree of construct under-representation. Third, the FOI components were not equally represented across each of the TSI-A instruments. While Campbell and Fiske (1959) acknowledged that an MTMM matrix could have an unbalanced design, they also stressed the importance of being able to compare multiple components across multiple methods. In addition, the FOI components of adherence, quality, or exposure were not as comprehensively addressed as participant responsiveness on the instruments I used. Although this did provide more insight into an underreported FOI component, I was not able to fully examine each of the components to the extent that I had hoped. Together, the unbalanced MTMM matrix and the underrepresentation of some of the FOI components restricted my ability to adequately address the convergent and discriminant validity of all the components.

APPENDIX A
FIDELITY OF IMPLEMENTATION SCALE DEVELOPMENT

Teacher Activity Reflection

Adherence Scale

A total of four items/item groups were selected for the reflection adherence scale. After initial internal consistency was calculated, one item was removed, and a total of three items/item groups were kept for the final Reflection adherence scale. In Table A.1, I present the item list, the initial and final correlations with the total score, and the internal consistency results for the Reflection adherence scale.

Quality Scale

A total of four items/item groups were selected for the reflection quality scale. After initial internal consistency was calculated, one item was removed, and a total of three items/item groups were kept for the final Reflection quality scale. In Table A.2, I present the item list, the

Table A.1
Internal Consistency Results for the Items That Comprised the Reflection Adherence Scale

Item label		Initial correlation with total	Final correlation with total
To what extent did you have your students follow the Exploring Our Fluid Earth (EOFE) procedures for this activity? [a]		.17	-
To what extent did you connect the activity to the ocean?		.58	.61
To what extent did you explicitly address the components of TSI with your students during this activity?		.64	.61
To what extent did you explicitly address the components of TSI with your students during this activity?	Phases of Inquiry	.54	.58
	Modes of Inquiry		
	Demeanors of scientist		
	Practices of scientists		
	Metacognition		
Reflection adherence scale's internal consistency		.69	.76

[a]Item was removed from final Reflection adherence scale

Table A.2
Internal Consistency Results for the Items That Comprised the Reflection Quality Scale

Item label		Initial correlation with total	Final correlation with total
How well do you think you guided your students through the TSI Phases of Inquiry for this activity?	Initiation	.75	.83
	Invention		
	Investigation		
	Interpretation		
	Instruction		
How well do you think you guided your students through the TSI Modes of Inquiry for this activity?	Curiosity	.58	.66
	Description		
	Authoritative Knowledge		
	Experimentation		
	Product evaluation		
	Technology		
	Replication		
	Induction		
	Deduction		
	Transitive knowledge		
How well do you think you used good questioning strategies for this activity?[a]		.39	-
How well do you think you implemented your assessment strategies for this activity?		.65	.62
Reflection quality scale's internal consistency		.77	.83

[a]Item was removed from final Reflection quality scale

initial and final correlations with the total score, and the internal consistency results for the Reflection quality scale.

Participant Responsiveness Scale

A total of nine items/item groups were selected for the reflection participant responsiveness scale. After initial internal consistency was calculated, two items were removed. A total of seven items/item groups were kept for the final Reflection participant responsiveness scale. In Table A.3, I present the item list, the initial and final correlations with the total score, and the internal consistency results for the Reflection participant responsiveness scale.

Table A.3
Internal Consistency Results for the Items That Comprised the Reflection Participant Responsiveness Scale

Item label	Initial correlation with total	Final correlation with total
To what extent do you think connecting the activity to the ocean helped engage your students?[a]	.35	-
Overall, how much do you think this activity helped improve your students' knowledge of the process of science?[a]	.43	-
Overall, how much do you think this activity helped improve your students' science content knowledge?	.79	.74
How successful do you think the process of planning using TSI was in improving your understanding of TSI?	.76	.76
Overall, how successful do you think you were in carrying out your planned TSI inquiry questioning strategies for this activity?	.77	.84
Overall, how useful do you think your questioning strategies were in helping you guide your students through the TSI phases and assess their progress?	.77	.78
To what extent has implementing this activity enhanced your understanding of teaching science as inquiry?	.81	.77
What is your level of understanding of this activity's content now that you have implemented the activity?	.53	.61
How confident are you with teaching this content now that you have implemented the activity?	.46	.54
Reflection participant responsiveness scale's internal consistency	.87	.93

[a]Item was removed from final Reflection participant responsiveness scale

Post-Cohort Questionnaire

Adherence Scale

A total of four items were selected for the PCQ adherence scale. After initial internal consistency was calculated, all four items were kept for the final PCQ adherence scale. In Table A.4, I present the item list, the final correlations with the total score, and the internal consistency results for the PCQ adherence scale.

Table A.4
Internal Consistency Results for the Items That Comprised the Post-Cohort Questionnaire Adherence Scale

Item label	Final correlation with total
To what extent did you implement all the steps of the target activities in each of the following modules? Module 1 (physical)	.80
To what extent did you implement all the steps of the target activities in each of the following modules? Module 2 (chemical)	.77
To what extent did you implement all the steps of the target activities in each of the following modules? Module 3 (biological)	.84
To what extent did you implement all the steps of the target activities in each of the following modules? Module 4 (ecological)	.76
PCQ adherence scale's internal consistency	.91

Exposure Scale

A total of one item group was selected for the PCQ exposure scale. After initial internal consistency was calculated, two sub-items from the item group were removed and eight sub-items from the item group were kept for the final PCQ exposure scale. In Table A.5, I present the item list, the initial and final correlations with the total score, and the internal consistency results for the PCQ exposure scale.

Table A.5
Internal Consistency Results for the Items That Comprised the Post-Cohort Questionnaire Exposure Scale

Item label		Initial correlation with total	Final correlation with total
How often did you include this mode in your instruction with your focus class?	Curiosity	.66	.63
	Description	.55	.56
	Authoritative Knowledge [a]	.39	-
	Experimentation [a]	.30	-
	Product Evaluation	.71	.68
	Technology	.87	.88
	Replication	.78	.76
	Induction	.68	.69
	Deduction	.71	.72
	Transitive Knowledge	.65	.65
Post-cohort questionnaire exposure scale's internal consistency		.87	.90

[a]Item was removed from final PCQ exposure quality scale

Participant Responsiveness Scale

A total of 13 items/item groups were selected for the PCQ participant responsiveness scale. After initial internal consistency was calculated, all 13 items/item groups were kept for the final PCQ participant responsiveness scale. In Table A.6, I present the item list, the final correlations with the total score, and the internal consistency results for the PCQ participant responsiveness scale.

Table A.6
Internal Consistency Results for the Items That Comprised the Post-Cohort Questionnaire Participant Responsiveness Scale

Item label		Final correlation with total
The target activities of [module no.] were successful in my focus class	Module1 (physical)	.69
	Module2 (chemical)	
	Module3 (biological)	
	Module4 (ecological)	
The target activities of [module no.] helped me to teach inquiry-based science	Module1 (physical)	.69
	Module2 (chemical)	
	Module3 (biological)	
	Module4 (ecological)	
How comfortable were you in implementing each of the following TSI activities with your focus class?	Teaching aquatic science content	.56
	Implementing the Module1 (physical) target activities	
	Implementing the Module2 (chemical) target activities	
	Implementing the Module3 (biological) target activities	
	Implementing the Module4 (ecological) target activities	
	Using a variety of modes of inquiry	
	Guiding students through the phases of inquiry	

Table A.6 (continued)
Internal Consistency Results for the Items That Comprised the Post-Cohort Questionnaire Participant Responsiveness Scale

Item label		Final correlation with total
To what degree do you value this mode in your instruction?	Curiosity	.64
	Description	
	Authoritative knowledge	
	Experimentation	
	Product Evaluation	
	Technology	
	Replication	
	Induction	
	Deduction	
	Transitive knowledge	
How comfortable are you in using each of the following modes in your instruction now that you have completed all four modules?	Curiosity	.67
	Description	
	Authoritative knowledge	
	Experimentation	
	Product Evaluation	
	Technology	
	Replication	
	Induction	
	Deduction	
	Transitive knowledge	
To what extent were each of the following TSI toolbox components useful in helping you teach science as inquiry?	Metacognition	.52
	Themes	
	Science as a discipline	
	Scientific demeanors	
	Practices of scientists	
	Questioning strategies	
	Teacher as research director	
How often will you use each of the following TSI toolbox components with your future students?	Metacognition	.56
	Themes	
	Science as a discipline	
	Scientific demeanors	
	Practices of scientists	
	Questioning strategies	
	Teacher as research director	
The TSI approach to teaching science as inquiry was valuable for my teaching practice		.83
I will continue to use what I learned in the TSI PD series		.66

Table A.6 (continued)
Internal Consistency Results for the Items That Comprised the Post-Cohort Questionnaire Participant Responsiveness Scale

Item label	Final correlation with total
The TSI PD series provided science content that was relevant to my teaching context	.54
The TSI pedagogical approaches were relevant to my teaching context	.76
The target activities engaged the students in my focus class	.57
The TSI PD series met my expectations of learning how to teach inquiry-based science	.60
Post-cohort questionnaire participant responsiveness scale's internal consistency	.86

Teacher Interview

Participant Responsiveness Scale

A total of four items were selected for the Interview participant responsiveness scale. After initial internal consistency was calculated, all four items were kept for the final participant responsiveness scale. In Table A.7, I present the item list, the final correlations with the total score, and the internal consistency results for the Interview participant responsiveness scale.

Table A.7
Internal Consistency Results for the Items That Comprised the Interview Participant Responsiveness Scale

Item label	Final correlation with total
How would you rate how valuable the TSI PD has been to your teaching practice?	.62
How would you rate how relevant the TSI PD has been to your teaching practice?	.74
How would you rate how successful you were in implementing the TSI content?	.57
How would you rate how successful you were in implementing the TSI pedagogy?	.71
Interview participant responsiveness scale's internal consistency	.84

Inquiring Into Science Instruction Observation Protocol (ISIOP)

Exposure Scale

A total of four items were selected for the ISIOP exposure scale. After initial internal consistency was calculated, one item was removed and three items were kept for the final exposure scale. In Table A.8, I present the item list, the initial and final correlations with the total ISIOP score, and the internal consistency results for the ISIOP exposure scale.

Table A.8
Internal Consistency Results for the Items That Comprised the ISIOP Exposure Scale

Item label	Initial correlation with total	Final correlation with total
How much of the instructional time was spent on questioning/exploration activities?	.48	.48
How much of the instructional time was spent on design activities?	.49	.49
How much of the instructional time was spent on data collection and organization activities? [a]	.30	-
How much of the instructional time was spent on analysis and conclusion activities?	.65	.65
ISIOP exposure scale's internal consistency	.67	.69

[a]Item was removed from final ISIOP exposure scale

Quality Scale

A total of 22 items were selected for the ISIOP quality scale. After initial internal consistency was calculated, 6 items were removed and 16 items were kept for the final ISIOP quality scale. In Table A.9, I present the item list, the initial and final correlations with the total ISIOP score, and the internal consistency results for the ISIOP quality scale.

Table A.9

Internal Consistency Results for the Items That Comprised the ISIOP Quality Scale

Item label	Initial correlation with total	Final correlation with total
The teacher facilitated a learning-conducive physical environment for the majority of the students.	.63	.65
The teacher projected a welcoming and engaging teaching style.	.45	.53
The teacher utilized teaching approaches to push students' thinking farther and encourage flexibility in their thinking.	.81	.75
The teacher stated the learning goals (i.e., the science content students would learn). [a]	.10	-
The teacher provided an overview of the activities in the lesson.[a]	< - .01	-
The teacher stated the performance expectations for the lesson (e.g., products, time frame).[a]	.04	-
The teacher situated the lesson within the context of previous lessons' science content.	.64	.65
The teacher clearly and explicitly connected the lesson's key science ideas to one another.	.70	.66
The teacher encouraged students to work together to develop collective understandings.	.46	.54
The teacher used adequate wait time (5 seconds or more) to allow students to formulate a response to questions.	.46	.63
The teacher encouraged students to respond to their classmates' thoughts and questions.	.85	.75
Transitions into the lesson and/or between lesson events were short in duration and did not interrupt instructional flow.	.88	.94
The teacher actively monitored individual and group progress (e.g., walking around the room to look at student work, asking for student verbal updates).	.72	.72
The teacher encouraged students to take responsibility for their learning by allowing them to make decisions about some aspect(s) of the class activity.	.70	.70
The teacher used formative assessment strategies to responsively pace the lesson.	.52	.57
The teacher facilitated student self-pacing of learning activities, when appropriate.	.51	.62
The teacher exhibited enthusiasm, curiosity, and interest in science.	.80	.80
The teachers' discourse and comments utilized students' thoughts, ideas, opinions, or questions as contributions to the class learning experience.	.76	.67
The teacher solicited from students what they know or believe about a topic in order to understand their prior conceptions.	.50	-

Table A.9 (continued)

Internal Consistency Results for the Items That Comprised the ISIOP Quality Scale

Students asked irrelevant questions of the teacher (e.g., personal, opinion, non-science or non-lesson related).[a]	-.51	-
The teacher asked students to expand on or clarify an idea previously offered by themselves, a peer, or other source of information.	.49	.51
The teacher exhibited openness to new ideas, approaches, and/or data.[a]	.44	-
ISIOP quality scale's internal consistency	.88	.93

[a]Item was removed from the final ISIOP quality scale.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

August, G. J., Egan, E. A., Realmuto, G. M., & Hektner, J. M. (2003). Parceling component effects of a multifaceted prevention program for disruptive elementary school children. *Journal of Abnormal Child Psychology*, *31*, 515-527.

Azano, A., Missett, T. C., Callahan, C. M., Oh, S., Brunner, M., Foster, L. H., & Moon, T. R. (2011). Exploring the relationship between fidelity of implementation and academic achievement in a third-grade gifted curriculum: A mixed-methods study. *Journal of Advanced Academics*, *22*, 693-719.

Backer, T. E. (2001). Finding the balance: Program fidelity and adaptation in substance abuse prevention: A state-of-the-art review. Rockville, MD: Center for Substance Abuse Prevention.

Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the pieces together: An integrated model of program implementation. *Prevention Science*, *12*(1), 23-33.

Berman, P., & McLaughlin, M. W. (1976, March). Implementation of educational innovation. In *The educational forum* (Vol. 40, No. 3, pp. 345-370). Taylor & Francis Group.

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, N., Davidson, W. S., Roitman, D. B., & Emshoff, J. G. (1987). The fidelity-adaptation debate: Implications for the implementation of public sector social programs. *American Journal of Community Psychology*, *15*, 253-268.

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H. W. (2000). Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Research, 2,* 75-87.

Botvin, G. J., Dusenbury, L., Baker, E., James-Ortiz, S., Botvin, E. M., & Kerner, J. (1992). Smoking prevention among urban minority youth: assessing effects on outcome and mediating variables. *Health Psychology*, *11*, 290-299.

Brandon, P. R., Taum, A. K., Ayala, C.C., Young, D. B., Gray M. E., Speitel, T. W., Nguyen, T. T., & Pottenger III, F. M. (2007). *Phase-I study of the effects of professional development and long-term support on program implementation and scaling up: final report*. Honolulu: University of Hawai'i at Mānoa, Curriculum Research & Development Group. http://manoa.hawaii.edu/crdg/wp-content/uploads/SCUP_final_report.pdf

Brandon, P. R., Taum, A. K., Young, D. B., Pottenger III, F. M., & Speitel, T. W. (2008). The complexity of measuring the quality of program implementation with observations: The case of middle school inquiry-based science. *American Journal of Evaluation*, *29*, 235-250.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*(1): 40. doi: 10.1186/1748-5908-2-40.

Century, J., & Cassata, A. (2016). Implementation Research: Finding Common Ground on What, How, Why, Where, and Who. *Review of Research in Education*, *40*, 169-215.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*, 199-218.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309-319.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago: Rand McNally.

Cronbach, L. J. (1971). *Test validation*. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23-45.

Darrow, C. L. (2013). The effectiveness and precision of intervention fidelity measures in preschool intervention research. *Early Education & Development*, *24*, 1137-1160.

Dhillon, S., Darrow, C., & Meyers, C. V. (2015). Introduction to implementation fidelity. In C. Meyers, C, & W. C. Brandt (Eds.), *Implementation fidelity in education research: Designer and evaluator considerations* (pp. 8-22).New York: Routledge.

Dillman, D. A. (2007). *Mail and internet surveys: The tailored design method*. Hoboken, NJ: Wiley & Sons.

Domitrovich, C. E., & Greenberg, M. T. (2000). The study of implementation: Current findings from effective programs that prevent mental disorders in school-aged children. *Journal of Educational and Psychological Consultation*, *11*, 193-221.

Dumas, J. E., Lynch, A. M., Laughlin, J. E., Smith, E. P., & Prinz, R. J. (2001). Promoting intervention fidelity: Conceptual issues, methods, and preliminary results from the EARLY ALLIANCE prevention trial. *American Journal of Preventive Medicine*, *20*(1), 38-47.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327-350.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research*, *18*, 237-256.

Emshoff, J. G., Blakely, C., Gottschalk, R., Mayer, J., Davidson, W. S., & Erickson, S. (1987). Innovation in education and criminal justice: Measuring fidelity of implementation and program effectiveness. *Educational Evaluation and Policy Analysis*, *9*, 300-311.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature* (FMHI Publication No. 231) Tampa: University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network. Retrieved November 29, 2010.

Fraenkel, J. R., & Wallen, N. E. (2003). *How to design and evaluate research in education*. New York: McGraw-Hill.

Francis, J. E., & White, L. (2002). Pirqual: a scale for measuring customer expectations and perceptions of quality in internet retailing. *AMA Winter Educators' Conference. Vol*, *13*, *13*, 263-269.

Frank, J. L., Bose, B., & Schrobenhauser-Clonan, A. (2014). Effectiveness of a school-based yoga program on adolescent mental health, stress coping strategies, and attitudes toward

violence: findings from a high-risk sample. *Journal of Applied School Psychology*, *30*(1), 29-49.

Giles, S., Jackson-Newsom, J., Pankratz, M. M., Hansen, W. B., Ringwalt, C. L., & Dusenbury, L. (2008). Measuring quality of delivery in a substance use prevention program. *The Journal of Primary Prevention*, *29*, 489-501.

Gould, L. F., Dariotis, J. K., Greenberg, M. T., & Mendelson, T. (2016). Assessing fidelity of implementation (FOI) for school-based mindfulness and yoga interventions: a systematic review. *Mindfulness*, *7*(1), 5-33.

Greenberg, M. T., Domitrovich, C. E., Graczyk, P. A., & Zins, J. E. (2005). The study of implementation in school-based preventive interventions: Theory, research, and practice. *Promotion of Mental Health and Prevention of Mental and Behavioral Disorders 2005 Series V3*.

Gresham, F. M., Gansle, K. A., & Noell, G. H. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis*, *26*, 257-263.

Gresham, F. M., MacMillan, D. L., Beebe-Frankenberger, M. E., & Bocian, K. M. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research & Practice*, *15*, 198-205.

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching: Theory and Practice*, *8*, 381-391.

Hallfors, D., & Godette, D. (2002). Will the 'Principles of Effectiveness' improve prevention practice? Early findings from a diffusion study. *Health Education Research*, *17*, 461-470.

Hansen, W. B., & McNeal, R. B. (1999). Drug education practice: Results of an observational study. *Health Education Research*, *14*(1), 85-97.

Henggeler, S. W., Schoenwald, S. K., Liao, J. G., Letourneau, E. J., & Edwards, D. L. (2002). Transporting efficious treatments to field settings: The link between supervisory practices and therapist fidelity in MST programs. *Journal of Clinical Child and Adolescent Psychology, 31*, 155–167.

Hernandez, M., Gomez, A., Lipien, L., Greenbaum, P. E., Armstrong, K. H., & Gonzalez, P. (2001). Use of the system-of-care practice review in the national evaluation: Evaluating the fidelity of practice to system-of-care principles. *Journal of Emotional and Behavioral Disorders*, *9*(1), 43-52.

Hill, L. G., & Owens, R. W. (2013). Component analysis of adherence in a family intervention. *Health Education, 113*, 264–280.

Ibrahim, S., & Sidani, S. (2016). Intervention Fidelity in Interventions: An Integrative Literature Review. *Research and Theory for Nursing Practice*, *30*, 258-271.

Kane, M. T. (2006). *Validation*. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17– 64). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology, 65*, 507-516.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Lane, K. L., Bocian, K. M., MacMillan, D. L., & Gresham, F. M. (2004). Treatment integrity: An essential—but often forgotten—component of school-based interventions. *Preventing School Failure: Alternative Education for Children and Youth*, *48*(3), 36-43.

LeLaurin, K., & Wolery, M. (1992). Research standards in early intervention: Defining, describing, and measuring the independent variable. *Journal of Early Intervention, 16,* 275-287.

Lucca, A. M. (2000). A Clubhouse Fidelity Index: Preliminary reliability and validity results. *Mental Health Services Research*, *2*, 89-94.

Lynch, S. (2007, April). A model for fidelity of implementation in a study of a science curriculum unit: Evaluation based on program theory. In *annual meeting of the National Association for Research in Science Teaching,* Chicago.

Lynch, S., & O'Donnell, C. (2005, April). The evolving definition, measurement, and conceptualization of fidelity of implementation in scale-up of highly rated science curriculum units in diverse middle schools. In *Annual Meeting of the American Educational Research Association*, Montreal.

Macias, C., Propst, R., Rodican, C., & Boyd, J. (2001). Strategic planning for ICCD clubhouse implementation: Development of the Clubhouse Research and Evaluation Screening Survey (CRESS). *Mental Health Services Research, 3,* 155-167.

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal, 21*, 341-366.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Metz, S. M., Frank, J. L., Reibel, D., Cantrell, T., Sanders, R., & Broderick, P. C. (2013). The effectiveness of the learning to BREATHE program on adolescent emotion regulation. *Research in Human Development, 10*, 252-272.

Meyers, C., & Brandt, W. C. (Eds.). (2014). *Implementation fidelity in education research: Designer and evaluator considerations*. New York: Routledge.

Mihalic, S. (2004). The importance of implementation fidelity. *Emotional and Behavioral Disorders in Youth*, *4*, 83-105.

Mills, S. C., & Ragan, T. J. (2000). A tool for analyzing implementation fidelity of an integrated learning system. *Educational Technology Research and Development*, *48*(4), 21-41.

Minner, D., & DeLisi, J. (2012). Inquiring into Science Instruction Observation Protocol (ISIOP). Waltham, MA: Education Development Center.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, *47*, 474-496.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, *11*, 247-266.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*, 315-340.

Muis, K. R., Winne, P. H., & Jamieson-Noel, D. (2007). Using a multitrait-multimethod analysis to examine conceptual similarities of three self-regulated learning inventories. *British Journal of Educational Psychology*, *77*, 177-195.

Noell, G. H., Gresham, F. M., & Gansle, K. A. (2002). Does treatment integrity matter? A preliminary investigation of instructional implementation and mathematics performance. *Journal of Behavioral Education*, *11*(1), 51-67.

Noell, G. H., Witt, J. C., Gilbertson, D. N., Ranier, D. D., & Freeland, J. T. (1997). Increasing teacher intervention implementation in general education settings through consultation and performance feedback. *School Psychology Quarterly, 12*, 77-88.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychological theory*. New York: MacGraw-Hill.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research, 78*(1), 33-84.

Pankratz, M. M., Jackson-Newsom, J., Giles, S. M., Ringwalt, C. L., Bliss, K., & Bell, M. L. (2006). Implementation fidelity in a teacher-led alcohol use prevention curriculum. *Journal of Drug Education, 36*, 317-333.

Resnick, B., Bellg, A. J., Borrelli, B., De Francesco, C., Breger, R., Hecht, J., ... & Ogedegbe, G. (2005). Examples of implementation and evaluation of treatment fidelity in the BCC studies: where we are and where we need to go. *Annals of Behavioral Medicine*, *29*(2), 46-54.

Resnicow, K., Davis, M., Smith, M., Lazarus-Yaroch, A., Baranowski, T., Baranowski, J., ... & Wang, D. T. (1998). How best to measure implementation of school health curricula: a comparison of three measures. *Health Education Research*, *13*, 239-250.

Ringwalt, C. L., Ennett, S., Vincus, A., Thorne, J., Rohrbach, L. A., & Simons-Rudolph, A. (2002). The prevalence of effective substance use prevention curricula in US middle schools. *Prevention Science*, *3*, 257-265.

Rogers, E. M. (1962). *Diffusion of innovations*. New York: Free Press of Glencoe.

Ross, J. A., McDougall, D., Hogaboam-Gray, A., & LeSage, A. (2003). A survey measuring elementary teachers' implementation of standards-based mathematics teaching. *Journal for Research in Mathematics Education, 34*, 344-363.

Ruiz-Primo, M. A. (2006, February). *A multi-method and multi-source approach for studying fidelity of implementation* (CSE Report 677). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ryan, B., & Gross, N. C. (1943). The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology*, *8*, 665-708.

Salkind, N. J. (Ed.). (2010). *Encyclopedia of research design* (Vol. 1). Thousand Oaks, CA: Sage.

Sánchez, V., Steckler, A., Nitirat, P., Hallfors, D., Cho, H., & Brodish, P. (2007). Fidelity of implementation in a treatment effectiveness trial of reconnecting youth. *Health Education Research, 22*, 95-107.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, *10*(1), 1-22.

Seraphin, K. (Ed.) (2014). *Accessible professional development for teaching aquatic science inquiry: Final report*. Honolulu: University of Hawai‘i at Mānoa, Curriculum Research & Development Group. http://manoa.hawaii.edu/crdg/wp-content/uploads/TSI-CRDG-final-report-Secured-Final-Version.pdf

Teague, G. B., Bond, G. R., & Drake, R. E. (1998). Program fidelity in assertive community treatment: development and use of a measure. *American Journal of Orthopsychiatry, 68*, 216-232.

Teague, G., Drake, R., & Ackerson, T. (1997). Evaluating Use of Continuous Treatment Teams for Persons With Mental Illness and Substance Abuse. *Year Book of Psychiatry and Applied Mental Health, 1997*, 191-192.

Weisman, A., Tompson, M. C., Okazaki, S., Gregory, J., Goldstein, M. J., Rea, M., & Miklowitz, D. J. (2002). Clinicians' Fidelity to a Manual-Based Family Treatment as a Predictor of the One-Year Course of Bipolar Disorder. *Family Process, 41*, 123-131.

Werts, C. E., & Linn, R. L. (1970). Cautions in applying various procedures for determining the reliability and validity of multiple-item scales. *American Sociological Review, 35*, 757-759.

Wilder, D. A., Atwell, J., & Wine, B. (2006). The effects of varying levels of treatment integrity on child compliance during treatment with a three-step prompting procedure. *Journal of Applied Behavior Analysis, 39*, 369-373.

Zvoch, K. (2012). How does fidelity of implementation matter? Using multilevel models to detect relationships between participant outcomes and the delivery and receipt of treatment. *American Journal of Evaluation, 33*, 547-565.